Peak Age of Information Distribution for Edge Computing With Wireless Links

Federico Chiariotti, *Member, IEEE*, Olga Vikhrova, Beatriz Soret, *Member, IEEE*,
and Petar Popovski, *Fellow, IEEE*

***Abstract*—Age of Information (AoI) is a critical metric for several Internet of Things (IoT) applications, where sensors keep track of the environment by sending updates that need to be as fresh as possible. The development of edge computing solutions has moved the monitoring process closer to the sensor, reducing the communication delays, but the processing time of the edge node needs to be taken into account. Furthermore, a reliable system design in terms of freshness requires the knowledge of the full distribution of the Peak AoI (PAoI), from which the probability of occurrence of rare, but extremely damaging events can be obtained. In this work, we model the communication and computation delay of such a system as two First Come First Serve (FCFS) queues in tandem, analytically deriving the full distribution of the PAoI for the $M/M/1$ – $M/D/1$ and the $M/M/1$ – $M/M/1$ tandems, which can represent a wide variety of realistic scenarios.**

***Index Terms*—Age of information, peak age of information, edge computing, queuing networks.**

I. Introduction

TRADITIONAL communication networks consider packet delay as the one and only performance metric to capture the latency requirements of a transmission. However, numerous IoT applications require the transmission of real-time status updates of a process from a generating point to a remote destination [1]. Sensor networks, vehicular networks and other tracking systems, and industrial control are examples of this kind of update process. For these cases, the AoI is a novel concept that better represents timeliness requirements by quantifying the freshness of the information at the receiver [2]. Basically, AoI computes the time elapsed since the latest received update was generated at any given moment in time, i.e., how old is the last packet received by the destination. Another age-related metric is the PAoI, which is the maximum value of AoI for each update, i.e., how old the last packet was when the next one is received by the destination. As in other performance metrics of communication systems, the PAoI is more informative than the average age when the interest is in worst-case analysis, e.g., when the system requirement is on the tail of the distribution. To illustrate, the PAoI can be useful to limit the latency of networked control systems, ensuring that the receiver has a recent picture of the state of the transmitter.

Edge computing is a technology that is gaining traction in age-sensitive IoT applications, as the transmission of sensor readings to a centralized cloud requires too much time and increases uncertainty, while processing the received data closer to the sensor that generated them can reduce the communication latency and the overall age of the information available to the monitoring or control process [3]. Freshness of information is going to be one of the critical parameters in beyond 5G and 6G, enabling Communications, Computing, Control, Localization, and Sensing (3CLS) services [4]: these applications require joint sensing, computation, and communication resources. In particular, services in telehealth, agriculture, manufacturing plants and robotics require strict control performance guarantees, which can only be possible by carefully designing the communication system. Furthermore, AoI and PAoI are often the relevant timing metrics for control systems, as they represent the time difference between the system observed by the controller and the real one when the control action takes place. As these services require high reliability, analyzing the average age is not enough: the tail of its distribution is also a very important parameter, as it directly affects the risk of control system failures, which must be constrained to very low levels for critical applications. In these edge applications that combine communication and computation, the contribution of both to the AoI needs to be taken into account. Limiting the age of the processed information is a critical requirement, which can influence the choice between local and edge-based computation [5] for IoT nodes. The problem becomes even more complex when considering multiple sources and different packet generation behaviors, along with limited communication capabilities [6].

Tandem queues (specifically the minimal 2-nodes case) are then a natural modeling choice for this kind of scenarios, as the first system represents the communication link, while the second one represents the computing-enabled edge node and its task queue. Fig. 1 shows an example: the communication buffer at the sensor and the task queue on the edge

Manuscript received October 2, 2020; revised December 7, 2020; accepted January 13, 2021. Date of publication January 20, 2021; date of current version May 18, 2021. This work has been in part supported by the Danish Council for Independent Research, Grant 8022-00284B SEMIOTIC, and by the European Union's Horizon 2020 research and innovation program as part of the IntellIoT project, which received funding under grant agreement 957218. The associate editor coordinating the review of this article and approving it for publication was T. He. *(Corresponding author: Federico Chiariotti.)*

Federico Chiariotti, Beatriz Soret, and Petar Popovski are with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: fchi@es.aau.dk).

Olga Vikhrova is with the DIIES Department, University Mediterranea of Reggio Calabria, 89124 Reggio Calabria, Italy.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2021.3053038.

Digital Object Identifier 10.1109/TCOMM.2021.3053038

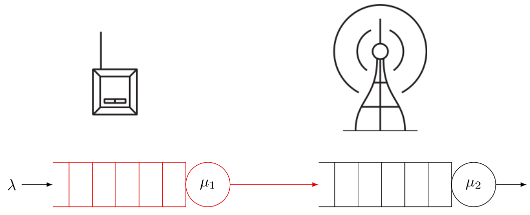


Fig. 1. An example of the IoT edge computing use case: the sensor transmits data to a computing-enabled edge node, which needs to maintain information freshness.

computing-enabled Base Station (BS) are the two queues in the tandem.

If the load on the computing-enabled edge node is time constant, while communication is less predictable due to, e.g., dynamic channel variations and random access, then the $M/M/1 - M/D/1$ tandem queue is an appropriate model. An $M/M/1$ queue is a proper model for the channel if we assume an ALOHA system [7] with perfect Multi-Packet Reception (MPR), in which [8] the packets are not lost due to collisions and can only be lost due to channel errors. This model is suitable for IoT systems based on Ultra-Narrowband (UNB) transmissions, such as SigFox [9]. On the other hand, computation time is often modeled as a linear function of the data size in the literature [10], and updates of the same size would have a constant and deterministic service time, therefore well-represented by an $M/D/1$ queue. Another option is to consider the computing load as also variable over time, leading to stochastic computing time, and represent the two systems by $M/M/1$ queues with different service rates [11].

An example of the AoI dynamics is plotted in Fig. 2: the AoI grows linearly over time, then decreases instantly when a new update arrives. Naturally, the age at the destination is never lower than the age at the intermediate node, as each packet that reaches the destination has already passed through the intermediate node, but the dynamics between the two are not trivial and serve as a motivation for this work. We analyze the distribution of the PAoI in a tandem queue with two systems with independent service times and a single source, where each infinite queue follows the FCFS policy. We consider both the $M/M/1 - M/D/1$ tandem and the $M/M/1 - M/M/1$ case, covering common communication relaying and communication and computation scenarios. Our aim is to derive the complete Probability Density Function (PDF) of the PAoI for systems with arbitrary packet generation and service rates, allowing system designers to define reliability requirements using PAoI thresholds and derive the network specifications needed to meet those requirements.

Although the motivation for our analysis is IoT edge computing, we notice there are other age-sensitive applications that can use the same models and results of this paper. For instance, a tandem queue can model a relay networking in which a packet is transmitted through one or more buffer-aided intermediate nodes between transmitter and receiver to, e.g., overcome the physical distance between the two end-points. A good example is a satellite relay that connects the ground and a satellite (or vice versa) through another satellite. In this case, the links are highly unpredictable and depend on different factors such as positioning jitter [12] and conditions in the upper atmosphere. A tandem queue in which the servers represent the transmission over successive links can represent this kind of system, provided that the service (transmission) times of different links are independent. Blockchain is another application where AoI is a critical metric to validate transactions in real time, particularly when combining the use of the distributed ledger with IoT applications [13]. As nodes need to transmit information, which is then validated, the tandem model is a useful abstraction for the communication and computation time [14].

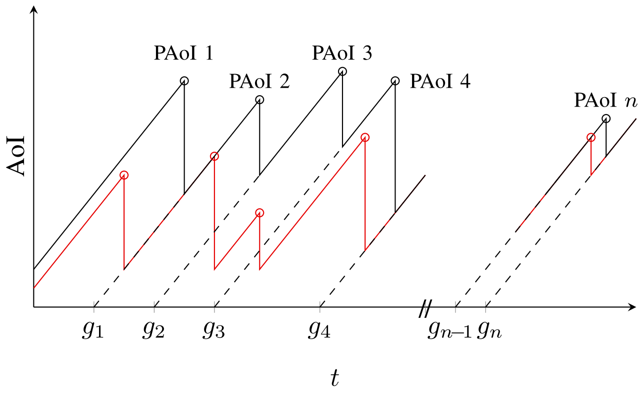


Fig. 2. The time-evolution of the Age of Information in a 2-node tandem queue. The age at the destination is plotted in black, the age at the intermediate node is plotted in red. Departures from each node are marked by a circle.

The contribution of this paper can be summarized by the following points:

- The full PDF of the PAoI is derived for the tandem of an $M/M/1$ and an $M/D/1$ queue, which is relevant for IoT edge computing scenarios;
- The same derivation is performed for the tandem of two $M/M/1$ queues, which can be relevant in relay-based communications with two independent links;
- Design considerations are drawn for the two systems, using the analytical formulas as a basis for system optimization.

The structure of the paper is as follows. In Sec. III the system model is detailed, as well as the procedure to calculate the AoI. Sec. V presents the calculations based on the model for the $M/M/1 - M/M/1$ tandem, while Sec. IV does the same for the $M/M/1 - M/D/1$ tandem. Numerical results are plotted in Sec. VI, and the paper is concluded in Sec.VII.[1]

II. Related Work

An overview of edge computing in IoT can be found in [3] and [15]. The initial research approach to latency in the edge paradigm was, in the context of 5G systems, to understand its potential to support Ultra Reliable Low Latency Communication (URLLC) [16]–[18]. The relevance of the AoI for edge computing applications was first identified in [5], although only the average AoI is computed. The same authors propose in [19] a joint transmission and computing scheduling

[1] The code used in the simulations, with the implementation of all the theoretical derivations in the paper, is available at https://github.com/AAU-CNT/tandem_aoi.

for a deadline. Another research area is the use of machine learning, particularly deep learning techniques, to unleash the full potential of IoT edge computing and enable a wider range of application scenarios [20], [21].

AoI is a relatively new metric in networking, but it has gained widespread recognition thanks to its relevance to several applications. Most theoretical results refer to simple queuing systems with a single node and FCFS policy. However, a few recent works focus their attention in the study of the age in tandem queues, even with different policies such as Last Come First Serve (LCFS) [22]. Some IoT scenarios have also been modeled as tandem $M/M/1$, $M/D/1$, or $D/M/1$ queues with multiple sources, as the read from the sensor is first pre-processed, and then transmitted to the server [23]. In this case, each queue follows the FCFS discipline, but the authors derive only the average PAoI. Another possibility is queue replacement, in which only the freshest update for each source is kept in the queue, reducing queue size significantly: in this case, the replaced packet is not placed in the queue, but dropped altogether, reducing channel usage with respect to simple LCFS, with or without preemption. In this case, the queue is modeled as an $M/M/1/2$, and if a new packet arrives it takes the queued packet's place. Some preliminary results on such a system are given in [24], while the average AoI and PAoI are computed in [25] for one source and in [26] for multiple sources. An analysis of the effect of preemption on this kind of models on the average AoI is presented in [27], and [28] derives the average AoI for two queues in tandem with preemption and different arrival processes. Another work [29] derives the full distribution of the PAoI for preemptive queues whose service time follows a phase-type distribution, which can be used to represent multiple hops. The PDF of the AoI in multi-hop networks with packet preemption has been derived in [30]. Another work [31] deals with multicast networks in which a single source updates multiple receivers over several hops, deriving the average age in that context. Another recent work considers a scenario in which packets that are not received within a deadline are dropped, deriving the average PAoI [32]. Some of these works deal with multi-hop queuing networks, of which our tandem model is a specific case, but they all assume some form of preemption. The computation of AoI and PAoI in systems with preemption is much simpler, particularly in systems of $M/M/1$ queues, but it cannot represent all the relevant use cases: preemption might not be possible or desirable, depending on the specific requirements of the control or monitoring application. For example, telehealth applications might benefit from receiving even out of date samples. To the best of our knowledge, this work is the first to derive the complete PDF of the age for non-preemptive tandem networks, covering these applications. Another example is the satellite relay scenario, where the traffic entering the tandem system comes from an aggregation of devices and therefore each individual update is relevant.

Other works concentrate on more realistic models, considering the effect of physical and medium access issues on the AoI. A model considering a fading wireless channel with retransmissions was used to compute the PAoI distribution over a single-hop link in [33], and a recent live AoI measurement study on a public networks generally confirmed that the theoretical models are realistic [34]. Other works compute the average AoI in Carrier Sense Multiple Access (CSMA) [35], ALOHA [36] and slotted ALOHA [37] networks, considering the impact of the different medium access policies on the age. It is also possible to jointly optimize the sampling and updating processes, i.e., both the reading instants from the sensor and the transmission of updates, if both are controllable [38]: the cost of both operations is a determinant of the overall AoI of the system [39].

A generalization of our model, relevant for applications like satellite relaying, is the multi-hop network with and arbitrary number of $M/M/1$ systems, senders, and receivers: in this case, the moment generating function of the AoI was derived in [40] for preemptive servers. Other scheduling policies make things more complicated: tight bounds for the average AoI were derived in [41] for the FCFS discipline and other AoI-oriented queue prioritization mechanisms.

Deriving the complete distribution of the age might be critical for reliability-oriented applications, but it is still mostly unexplored in the literature: while the work on deriving the first moments of the AoI distribution is extensive, the analytical complexity of deriving the complete PDF is a daunting obstacle. A recent work [42] uses the Chernoff bound to derive an upper bound of the quantile function of the AoI for two queues in tandem with deterministic arrivals, but to the best of our knowledge, the complete PAoI distribution has only been derived for simple queuing systems [33]. Another interesting approach to achieve reliability is to consider the AoI process directly: in [43], the authors use extreme value theory to derive the complementary Cumulative Density Function (CDF) of the PAoI in a realistic channel setting with scheduled transmissions. [44] proposes the use of the risk of ruin, an economic concept, as a metric of fresh and reliable information in augmented reality. Specifically, the CDF of the PAoI is used to find the probability of maximum severity of ruin PAoI in single node systems. Yet another approach to reliability, which does not derive the complete distribution of the AoI but uses an average measure on its tail, is presented in [45]: the authors pose the risk minimization problem as a Markov Decision Process (MDP), optimizing the age with bounded risk. For a more complete overview of the literature on AoI, we refer the reader to [46].

III. System Model

We consider a tandem system composed by two consecutive queues, where the first one is $M/M/1$. Packets are generated at the first system by a Poisson process with rate λ and enter the first queue, whose service time is exponentially distributed with rate μ_1. When a packet exits the first system, it enters the second one, whose service time is a constant D or an exponential random variable with rate μ_2. A simple diagram of the tandem is shown in Fig. 1. Both queues are of infinite size and are oblivious to the content of the packets: there is no preemption of the updates, i.e. an older packet is not removed from the queue when a new update comes from the same source. As explained in the introduction, we assume that

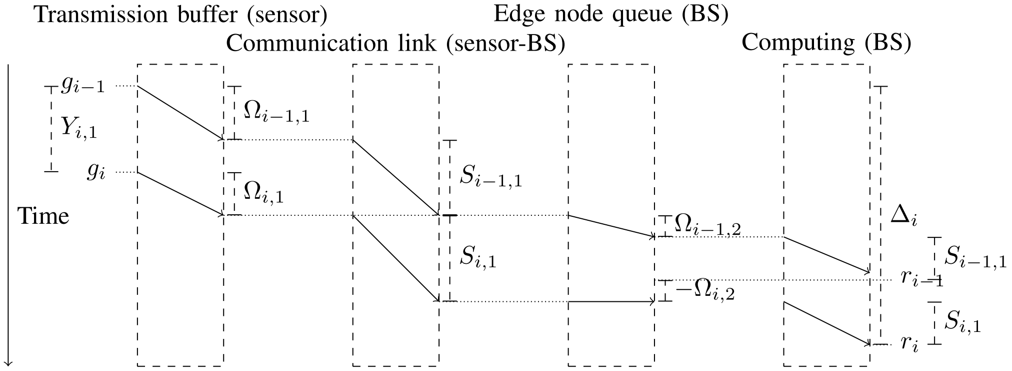


Fig. 3. Schematic of the four steps a packet goes through in a tandem queue, highlighting the components of the PAoI.

the service times in the two systems are independent. The assumption is realistic for edge computing systems, as the communication and processing are usually independent, but not always verified in relay networks; it therefore needs a careful examination. Even if the service times are independent, however, the waiting times are not, as the queue at the second system depends on the output of the first one. In the following, we use the compact notation $p_{X|Y}(x|y)$ for the conditioned probability $p[X = x|Y = y]$. PDFs are denoted by a lower-case p, and CDFs by an upper-case P.

In a tandem queue, the packet generation times correspond to the arrival times at the first queue, whereas the receiving instants are the departure times in the second queue. We define the total system time for packet i as T_i: when a packet is received, the AoI is equal to T_i, i.e., the difference between the time r_i when it is received by the destination and the time g_i when it was generated. The PAoI (see Fig. 2) is the maximum value of the AoI, i.e., the age at the instant immediately before the arrival of a new update. If we denote the interarrival time $Y_i = g_i - g_{i-1}$, then PAoI is given by

$$\Delta_i = r_i - g_{i-1} = r_i - g_i + g_i - g_{i-1} = T_i + Y_i. \quad (1)$$

If packet i arrives right after packet $i-1$, it will probably have to wait in the queue for it to depart the system: the system time T_i depends on the interarrival time Y_i. The PDF of the PAoI with value τ_i, denoted as $p_{\Delta_i}(\tau_i)$, can then be computed by using the conditional system time probability $p_{T_i|Y_i}(t_i|y_i)$:

$$p_{\Delta_i}(\tau_i) = \int_0^{\tau_i} p_{Y_i}(y_i) p_{T_i|Y_i}(\tau_i - y_i|y_i) dy_i. \quad (2)$$

We then need to compute $p_{T_i|Y_i}(t_i|y_i)$. For each system $j = 1, 2$ in the tandem, the system time $T_{i,j}$ is defined as the sum of the waiting time $W_{i,j}$ and the service time $S_{i,j}$. We also define $Y_{i,j}$, the interarrival time at system j. For $j = 1$, we have $Y_{i,1} = Y_i$, while for $j = 2$:

$$Y_{i,2} = g_i + T_{i,1} - (g_{i-1} + T_{i-1,1}) = Y_i + T_{i,1} - T_{i-1,1}. \quad (3)$$

Since the first queue is $M/M/1$, the system times for the two queues are independent, as proven by Reich [47] using Burke's theorem [48] and considering each system in steady state for packet $i-1$. If we consider system j in steady state, i.e., we do not condition on $Y_{i-1,j}$, the system time $T_{i-1,j}$ is exponentially distributed with rate $\alpha_j = \mu_j - \lambda$. However, the values of Y_i and T_i are correlated, and the computation of the PAoI needs to account for this fact. In the following, we give the conditional PDF of the components of the PAoI, which will then be joined in the derivation. 6 First, we define the *extended waiting time* $\Omega_{i,j}$ as the difference between the previous packet's system time and the interarrival time at the system, i.e., $\Omega_{i,j} = T_{i-1,j} - Y_{i,j}$. The reason we named $\Omega_{i,j}$ the extended waiting time is that $W_{i,j} = [\Omega_{i,j}]^+$, where $[x]^+$ is equal to x if it is positive and 0 if x is negative. From the definition of $\Omega_{i,j}$, we have:

$$\begin{aligned} Y_{i,2} &= Y_i + (W_{i,1} + S_{i,1}) - T_{i-1,1} \\ &= S_{i,1} + W_{i,1} - \Omega_{i,1} \\ &= S_{i,1} + [-\Omega_{i,1}]^+. \end{aligned} \quad (4)$$

since $W_{i,1} - \Omega_{i,1} = [\Omega_{i,1}]^+ - \Omega_{i,1} = [-\Omega_{i,1}]^+$. In the following paragraphs, we derive the PDF of the extended waiting time for the two system types that we are analyzing. Fig. 3 shows a possible realization of a packet's path through the tandem queue, highlighting the meaning of the extended waiting time: in the first system, in which packet i is queued, it corresponds to the waiting time, while in the second, in which the packet is not queued and enters service immediately, its negative value corresponds to the time between the departure of packet $i-1$ from the second system and the arrival of packet i at the same system. When $W = 0$, we have a negative extended waiting time, as packet i arrives after packet $i-1$ leaves the system. In general, we have $\Omega_{i,2} = T_{i-1,2} - Y_{i,2}$, and the system time for packet $i-1$ is $T_{i-1,2} = W_{i-1,2} + D$, while we know the interarrival time $Y_{i,2} = S_{i,1} + [-\Omega_{i,1}]^+$.

Theorem 1: The CDF of the waiting time for an $M/D/1$ queue with arrival rate λ and service time D is:

$$P_W(w) = (1 - \lambda D) \sum_{k=0}^{\left\lfloor \frac{w}{D} \right\rfloor} \frac{(-\lambda(w - kD))^k e^{\lambda(w-kD)}}{k!}. \quad (5)$$

Proof: See the derivation by Erlang [49]. ■

Corollary 1: We can find the queuing time PDF by deriving the CDF from (5)*:*

$$p_W(w) = (1-\lambda D)\left(\lambda e^{\lambda w} + \sum_{k=1}^{\left\lfloor \frac{w}{D} \right\rfloor} \frac{(-\lambda)^k (w-kD)^{k-1}}{k!} \right.$$
$$\left. \times e^{\lambda(w-kD)}(k+\lambda(w-kD))\right) \quad \forall w > 0;$$
$$p_W(0) = 1-\lambda D. \tag{6}$$

The CDF of the waiting time has a discontinuity, as the waiting time is exactly 0 with probability $1-\lambda D$, which corresponds to the probability of the packet finding an empty system and entering service immediately.

Corollary 2: As service time in an $M/D/1$ system is constant, we have the CDF of the total time in the system:

$$P_T(t) = P_W(t-D)u(t-D). \tag{7}$$

Corollary 3: The PDF of $\Omega_{i,2}$ conditioned on $S_{i,1}$ and $\Omega_{i,1}$ is given by:

$$\begin{aligned} &p_{\Omega_{i,2}|S_{i,1},\Omega_{i,1}}(\omega_{i,2}|s_{i,1},\omega_{i,1}) \\ &\quad = p(W_{i-1,2} = \omega_{i,2} - s_{i,1} - [-\omega_{i,1}]^+ - D) \\ &\quad = p_W(\omega_{i,2} + s_{i,1} + [-\omega_{i,1}]^+ - D). \end{aligned} \tag{8}$$

Fig. 4 shows this clearly: when $\Omega_{i,1}$ is negative, the extended queuing time $\Omega_{i,2}$ corresponds to $S_{i,1}+W_{i-1,2}+D-\Omega_{i,1}$, while when $\Omega_{i,1}$ is positive, packet i starts service immediately after packet $i-1$ leaves the first system, and we have $\Omega_{i,2} = S_{i,1}+W_{i-1,2}+D$. The waiting time in $M/D/1$ system is analytically derived, although it can give numeric problems for very large waiting times and high load [50]. If the application requires the computation of very large waiting times with $\lambda D \simeq 1$, we suggest the use of more numerically stable methods from the relevant literature.

Theorem 2: The PDF of the extended waiting time in a $M/M/1$ - $M/M/1$ tandem queue is given by:

$$p_{\Omega_{i,j}|Y_{i,j}}(\omega_{i,j}|y_{i,j}) = \alpha_j e^{-\alpha_j(\omega_{i,j}+y_{i,j})} u(\omega_{i,j}+y_{i,j}), \tag{9}$$

where $u(\cdot)$ is the step function.

Proof: Knowing the PDF of the system time $T_{i-1,j}$ from well-known results on $M/M/1$ queues, the interarrival time at the first relay $Y_{i,1}$ is exponentially distributed with rate λ, while in subsequent systems it is given by $Y_{i,2} = S_{i,1} + [-\Omega_{i,1}]^+$. ■

Corollary 4: We can combine (9) *with the definition of $Y_{i,2}$ to get*:

$$\begin{aligned} p_{\Omega_{i,2}|S_{i,1},\Omega_{i,1}}(\omega_{i,2}|s_{i,1},\omega_{i,1}) &= \alpha_1 e^{-\alpha_1\left(\omega_{i,2}+s_{i,1}[-\omega_{i,1}]^+\right)} \\ &\quad \times u(\omega_{i,2}+s_{i,1}+[-\omega_{i,1}]^+). \end{aligned} \tag{10}$$

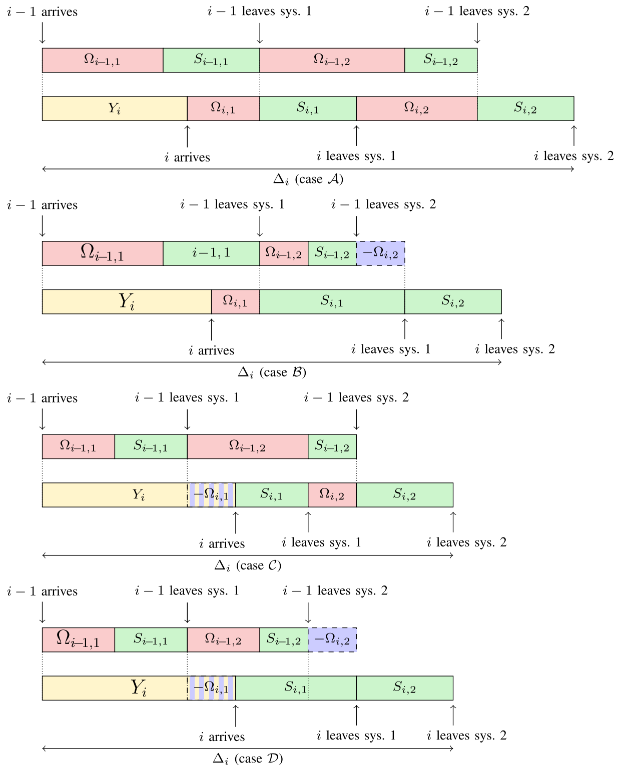


Fig. 4. Schematic of the components of the PAoI. (case $\mathcal{A}$): packet i has to wait in both queues. (case $\mathcal{B}$): packet i waits only in the first queue. (case $\mathcal{C}$): packet i waits only in the second queue. (case $\mathcal{D}$): packet i is immediately served in both systems. Observe the negative expected waiting time $\Omega_{i,j}$ in the cases with empty queue(s).

To compute the exact PDF of PAoI in the 2-system case ($j \in 1, 2$), we distinguish between free and busy systems at each node, i.e., we condition the PDF on the state of each system when packet i arrives to it and calculate it separately for the four possible combinations. Case $\mathcal{A}$ is defined as $\Omega_{i,1} > 0 \wedge \Omega_{i,2} > 0$, while in case $\mathcal{B}$ we have $\Omega_{i,1} > 0 \wedge \Omega_{i,2} \leq 0$. Similarly, in case $\mathcal{C}$ we have $\Omega_{i,1} \leq 0 \wedge \Omega_{i,2} > 0$ and in case $\mathcal{D}$ we get $\Omega_{i,1} \leq 0 \wedge \Omega_{i,2} \leq 0$. An example of the relevant values in the four cases are shown in Fig. 4: in case $\mathcal{A}$, packet i is queued in both systems, as the previous packet is still in the system when i arrives in each. The two extended queuing times (shown in red) are positive. In case $\mathcal{B}$, packet $i-1$ has already left the second system when packet i leaves the first: the extended queuing time (shown in blue with a dashed outline) is negative, and packet i enters service in the second system as soon as it arrives. In case $\mathcal{C}$, it is the first system that is empty when packet i arrives, and in case $\mathcal{D}$, both systems are empty, and the packet enters service directly at both.

We can then exploit Corollaries 1-3 to compute the conditioned PDF of the total time at the second system. The PAoI can then simply be computed by unconditioning over the values of $S_{i,1}$, $\Omega_{i,1}$, and Y_i, simply applying the law of total probability until the PDF for the given case is derived. The division in 4 cases is not strictly necessary, but it reduces the number of terms in the equations considerably with respect to deriving the PDF for the general case directly. The computation for the $M/M/1$ – $M/M/1$ tandem follows the same reasoning, using the results in Theorem 2 and its Corollary 4.

Definition 1: The PDF of the PAoI is:

$$\begin{aligned} p_\Delta(\tau) = &\, p_{\Delta|\mathcal{A}}(\tau)\,p(\mathcal{A}) + p_{\Delta|\mathcal{B}}(\tau)\,p(\mathcal{B}) \\ &+ p_{\Delta|\mathcal{C}}(\tau)\,p(\mathcal{C}) + p_{\Delta|\mathcal{D}}(\tau)\,p(\mathcal{D}). \end{aligned} \tag{11}$$

where $p_{\Delta|\mathcal{X}}$ is the PDF of the PAoI in case X and $p(\mathcal{X})$ is the probability of case X happening. The definition comes from the application of the total law of probability.

In case $\mathcal{A}$, packet i is queued at both systems, and the packet will have the highest queuing delay and system time. The case with the lowest system time is case $\mathcal{D}$, in which the packet experiences no queuing. However, these intuitive relations do not necessarily hold for the PAoI, as the interarrival time between update packets can play a major role. In the analysis of the four cases, we will omit the packet index i wherever possible for the sake of readability. For a quick overview of the notation used in the rest of this paper, we refer the reader to Table I.

IV. PAoI Distribution for the $M/M/1$ – $M/D/1$ Tandem

First, we analyze the tandem of an $M/M/1$ and an $M/D/1$ system, using the PDF of the waiting time from (6).

Definition 2: The auxiliary function $\theta(M,\beta)$, which we will use in the later derivations to make the notation more compact, is defined as:

$$\begin{aligned}
&\theta(M,\beta)\\
&= \int_0^M p_W(w)e^{\beta w}dw \,\forall\beta\neq 0, \beta\neq-\lambda\\
&= (1-\lambda D)\Bigg[\sum_{k=1}^{\lfloor\frac{M}{D}\rfloor}\int\limits_{kD}^{M}\frac{(-\lambda)^k(w-kD))^{k-1}}{k!}\\
&\qquad\times e^{\lambda(w-kD)+\beta w}(k+\lambda(w-kD))dw+\int\limits_0^M\lambda e^{(\lambda+\beta)w}dw\Bigg]\\
&= \frac{\lambda(1-\lambda D)}{\lambda+\beta}\Bigg(\sum_{k=1}^{\lfloor\frac{M}{D}\rfloor}\Bigg[\frac{\beta\lambda^{k-1}e^{\beta kD}}{(\lambda+\beta)^k}-\lambda^k e^{\lambda(M-kD)+\beta M}\\
&\qquad\times\left(\frac{(kD-M)^k}{k!}+\sum_{j=0}^{k-1}\frac{\beta(kD-M)^j}{\lambda(\lambda+\beta)^{k-j}j!}\right)\Bigg]\\
&\qquad+e^{(\lambda+\beta)M}-1\Bigg).
\end{aligned}\tag{12}$$

If $\beta=0$, the result of the integral is simply the waiting time CDF, i.e., $\theta(M,0)=P_W(M)$. If $\beta=-\lambda$, we have:

$$\begin{aligned}
\theta(M,-\lambda) &= \int_0^M p_W(w)e^{\beta w}dw\\
&= (1-\lambda D)\Bigg(\sum_{k=1}^{\lfloor\frac{M}{D}\rfloor}\int\limits_{kD}^{M}\frac{(-\lambda)^k(w-kD))^{k-1}e^{-\lambda kD}}{k!}\\
&\qquad\times(k+\lambda(w-kD))dw+\lambda M\Bigg)\\
&= (1-\lambda D)\Bigg(\lambda M+e^{\beta kD}(-\lambda)^k\\
&\qquad\times\left(\frac{(M-kD)^k}{k!}+\frac{\lambda(M-kD)^{k+1}}{(k+1)!}\right)\Bigg).
\end{aligned}\tag{13}$$

TABLE I

Main Notation Used in the Paper

Symbol	Description
λ	Packet generation rate
μ_j	Service rate of $M/M/1$ system j
D	Service time of the $M/D/1$ system
$\alpha_j=\mu_j-\lambda$	Response rate of system j
$S_{i,j}$	Service time in system j for packet i
$Y_{i,j}$	Interarrival time in system j for packet i
$\Omega_{i,j}=T_{i-1,j}-Y_{i,j}$	Extended waiting time in j for packet i
$W_{i,j}=[\Omega_{i,j}]^+$	Waiting time in system j for packet i
$T_{i,j}=S_{i,j}+W_{i,j}$	Total time in system j for packet i
$\Delta_i=Y_i+T_i$	PAoI for packet i

If we consider a simple $M/D/1$ queue (i.e., not in tandem), it is easy to derive the distribution of the PAoI, as we have

$$\Delta_i = D+\max(Y_i,T_{i-1}).\tag{14}$$

Therefore, the PAoI is lower than τ when both Y_i and T_{i-1} are smaller than $\tau-D$, and we can write the CDF as:

$$\begin{aligned}
P_\Delta(\tau) &= \int_0^{\tau-D}P_W(\tau-2D)\lambda e^{-\lambda y}dy\\
&= (1-e^{\lambda(T-\tau)})P_W(\tau-2D)u(\tau-2D).
\end{aligned}\tag{15}$$

Things are more complex in the tandem system, in which an $M/M/1$ queue feeds the $M/D/1$ queue. Thanks to Burke's theorem [48], we can consider both systems to be in steady state for packet $i-1$, distinguishing the same four cases $\mathcal{A}$-$\mathcal{D}$ described in Section III and Figure 4.

A. The Packet Is Queued at Both Systems

We start by considering the conditional CDF of the PAoI in case $\mathcal{A}$, in which the i-th packet is queued at both systems (i.e., $\Omega_{i,1}>0\wedge\Omega_{i,2}>0$). The probability of a packet being in case $\mathcal{A}$ is given by two concurrent events: first, packet i must find the first system busy, which is equivalent to stating that $T_{i-1,1}>Y_i$. Then, the second system must also be busy when the packet arrives to it, so we have $T_{i-1,1}+W_{i-1,2}+S_{i-1,2}>Y_i+S_{i,1}$:

$$\begin{aligned}
&p(\mathcal{A})\\
&\quad= p(\Omega_1>0)p(\Omega_2>0|\Omega_1>0)\\
&\quad= p(Y_i<T_{i-1,1},S_{i,1}<T_{i-1,1}-Y_i+W_{i-1,2}+D)\\
&\quad= \int\limits_0^\infty\int\limits_0^{t_1}p_{y_1}(y_1)p_{T_1}(t_1)dy_1dt_1\int\limits_0^\infty P_{S_1}(w+D)p_W(w)dw\\
&\quad= \rho_1\left(1-(1-\lambda D)e^{-\mu_1 D}-e^{-\mu_1 D}\lim_{M\to\infty}\theta(M,-\mu_1)\right)\\
&\quad= \rho_1\left(1-(1-\lambda D)e^{-\mu_1 D}\left(1+\frac{\lambda(e^{\mu_1 D}-1)}{\alpha_1 e^{\mu_1 D}+\lambda)}\right)\right).
\end{aligned}\tag{16}$$

The conditioned distribution of the PAoI in case $\mathcal{A}$ is:

$$\begin{aligned}
p_{\Delta_i|T_{i-1,1},\mathcal{A}}(\tau|t_1) &= \frac{p_W(\tau-t_1-2D)}{p(\mathcal{A})}\\
&\qquad\times P_{S_i}(\tau-D-t_1)P_{Y_i}(t_1)\\
&= \frac{(1-e^{-\lambda t_1})(1-e^{-\mu_1(\tau-t_1-D)})}{p(\mathcal{A})}\\
&\qquad\times p_W(\tau-t_1-2D)u(\tau-t_1-2D).
\end{aligned}\tag{17}$$

We need to consider the case for $w = 0$ separately, as there is a discontinuity in the CDF. We can now uncondition the distribution by substituting $p_W(w)$ and applying the law of total probability:

$$
\begin{aligned}
&p_{\Delta|\mathcal{A}}(\tau)\\
&= \int_0^{\tau-2D} p_{\Delta_i|T_{i-1,1},\mathcal{A}}(\tau|t_1)\alpha_1\ e^{-\alpha_1 t_1}dt_1\\
&\quad +\alpha_1(1-\lambda D)e^{-\alpha_1(\tau-2D)}(1-e^{-\lambda(\tau-2D)}-e^{-\mu_1 D})\\
&= \int_0^{\tau-2D} \frac{\alpha_1}{p(\mathcal{A})}\Big(e^{\alpha_1(w-\tau+2D)}-e^{\mu_1(w-\tau+2D)}\\
&\quad -e^{-\mu_1\ D-\alpha_1(\tau-2D)-\lambda w}+e^{-\mu_1(\tau-D)}\Big)p_W(w)dw\\
&\quad +\frac{\alpha_1(1-\lambda D)}{p(\mathcal{A})}(1-e^{-\lambda t})(1-e^{-\mu_1(\tau-t-D)})e^{-\alpha_1(\tau-2D)}.
\end{aligned}
\tag{18}
$$

We can then substitute the auxiliary function defined in Lemma 2 in (18), getting the PDF of the PAoI in case $\mathcal{A}$:

$$
\begin{aligned}
p_{\Delta|\mathcal{A}}(\tau) = \frac{\alpha_1}{p(\mathcal{A})}\Big(&e^{-\alpha_1(\tau+2D)}\theta(\tau-2D,\alpha_1)+e^{-\mu_1 D}\\
&\big(-e^{-\mu_1(\tau-3D)}\theta(\tau-2D,\mu_1)\\
&-e^{-\alpha_1(\tau-2D)}\theta(\tau-2D,-\lambda)\\
&+e^{-\mu_1(\tau-2D)}(P_W(\tau-2D)-(1-\lambda D))\big)\Big).
\end{aligned}
\tag{19}
$$

B. The Packet Is Only Queued at the First System

We now consider case $\mathcal{B}$, in which there is queuing at the first system, but not at the second. This is equivalent to stating that $T_{i-1,1} > Y_{i,1} \wedge \Omega_{i-1,2} + D \leq S_{i,1}$. Consequently, we get the following probability for case $\mathcal{B}$:

$$
\begin{aligned}
&p(\mathcal{B})\\
&= p(\Omega_1 > 0)p(\Omega_2 \leq 0|\Omega_1 > 0)\\
&= p(Y_i < T_{i-1,1}, S_{i,1} \geq T_{i-1,1} - Y_i + W_{i-1,2} + D)\\
&= \int_0^{\infty}\int_0^{t_1} p_{y_1}(y_1)p_{T_1}(t_1)dy_1dt_1\int_0^{\infty}(1-P_{S_1}(w+D))p_W(w)dw\\
&= \rho_1 e^{-\mu_1 D}\left((1-\lambda D)+\lim_{M\to\infty}\theta(M,-\mu_1)\right)\\
&= (1-\lambda D)\rho_1 e^{-\mu_1 D}\left(1+\frac{\lambda(e^{\mu_1 D}-1)}{\alpha_1 e^{\mu_1 D}+\lambda}\right).
\end{aligned}
\tag{20}
$$

The PAoI in this case is given by $T_{i-1,1} + S_{i,1} + D$. Its conditional PDF in this case is given by:

$$
\begin{aligned}
p_{\Delta|T_{i-1,1},\mathcal{B}}(\tau|t_1) &= \frac{p_{S_1}(\tau-t_1-D)}{p(\mathcal{B})}\\
&\quad \times P_{Y_1}(t_1)P_W(\tau-t_1-2D)\\
&= \frac{\mu_1\ e^{-\mu_1(\tau-t_1-D)}}{p(\mathcal{B})}\\
&\quad \times P_W(\tau-t_1-2D)(1-e^{-\lambda t_1}).
\end{aligned}
\tag{21}
$$

We can now uncondition this probability to obtain the PDF of the PAoI in case $\mathcal{B}$:

$$
\begin{aligned}
&p_{\Delta|\mathcal{B}}(\tau)\\
&\quad = \int_D^{\tau-D}\frac{P_W(s_1-D)}{p(\mathcal{B})}\alpha_1\ e^{-\alpha_1(\tau-s_1-D)}\\
&\qquad \times\mu_1\ e^{-\mu_1\ s_1}(1-e^{-\lambda(\tau-s_1-D)})dx\\
&\quad = \alpha_1\frac{\mu_1}{p(\mathcal{B})}e^{-\alpha_1(\tau-D)-\lambda D}\\
&\qquad \times\left(\int_0^{\tau-2D}P_W(w)e^{-\lambda w}dw-\int_0^{\tau-2D}P_W(w)dw\right)\\
&\quad = \frac{\alpha_1(1-\lambda D)}{\rho_1\ p(\mathcal{B})}e^{-\alpha_1(\tau-D)-\lambda D}\sum_{k=0}^{\lfloor\frac{\tau-2D}{D}\rfloor}\Bigg[e^{-\lambda(\tau-2D)}\\
&\qquad -\sum_{j=0}^{k+1}\frac{(-\lambda)^j(\tau-(k+2)D)^j e^{-\lambda kD}}{j!}\Bigg].
\end{aligned}
\tag{22}
$$

C. The Packet Is Only Queued at the Second System

We then consider case $\mathcal{C}$, in which there is no queuing at the first system, but the packet is queued at the second. This is equivalent to stating that $T_{i-1,1} \leq Y_{i,1} \wedge \Omega_{i-1,2} + D > S_{i,1} - \Omega_{i,1}$. Consequently, we get the following probability for case $\mathcal{C}$:

$$
\begin{aligned}
p(\mathcal{C}) &= p(\Omega_1 \leq 0, \Omega_2 > 0)\\
&= p(Y_i \geq T_{i-1,1}, S_{i,1} < T_{i-1,1} - Y_i + W_{i-1,2} + D)\\
&= \int_0^{\infty}\int_0^{y_1} p_{y_1}(y_1)p_{T_1}(t_1)\\
&\quad \times \int_{\max(0,y_1-t_1-D)}^{\infty} P_{S_1}(w+t_1+D-y_1)\\
&\quad \times p_W(w)dwdt_1\ dy_1 = \lambda D - p(\mathcal{A}).
\end{aligned}
\tag{23}
$$

The PAoI in this case is given by $T_{i-1,1} + \Omega_{i-1,2} + 2D$. Since we know that we are in case $\mathcal{C}$, we have:

$$
\begin{aligned}
&p_{\Delta|Y_{i,1},\Omega_{i-1,2},\mathcal{C}}(\tau|y_1,w)\\
&= \frac{P_{S_1}(\tau-y_1-D)P_{Y_1}(y_1)}{p(\mathcal{C})}\\
&\quad \times p_T(\tau-2D-w)u(y_1+w+2D-\tau)\\
&= \frac{\alpha_1(1-e^{-\mu_1(\tau-y_1-D)})e^{\alpha_1(w-\tau+2D)}u(y_1+w+2D-\tau)}{p(\mathcal{C})}.
\end{aligned}
\tag{24}
$$

We can now uncondition this probability on $Y_{i,1}$ by applying the law of total probability:

$$
\begin{aligned}
&p_{\Delta|\Omega_{i-1,2},\mathcal{C}}(\tau|w)\\
&\quad = \int_{\tau-2D-w}^{\tau-D}\frac{\lambda e^{-\lambda y_1}\alpha_1 e^{\alpha_1(w-\tau+2D)}}{p(\mathcal{C})}dy_1\\
&\qquad \times(1-e^{-\mu_1(\tau-y_1-D)})\\
&\quad = \frac{e^{\mu_1(D-\tau)}(\alpha_1 e^{\mu_1(w+D)}-\mu_1\ e^{\alpha_1(w+D)}+\lambda)}{p(\mathcal{C})}.
\end{aligned}
\tag{25}
$$

We can now uncondition again on $T_{i-1,i}$ and get the PDF of the PAoI in case $\mathcal{D}$:

$$\begin{aligned} &p_{\Delta|\mathcal{C}}(\tau)\\ &\quad=\int\limits_{0}^{\tau-2D} p_{\Delta|\Omega_{i-1,2},\mathcal{C}}(\tau|w)p_W(w)dw\\ &\qquad+\frac{1-\lambda D}{p(\mathcal{C})}p_{\Delta|\Omega_{i-1,2},\mathcal{C}}(\tau|0)\\ &\quad=\frac{e^{-\mu_1(\tau-D)}}{p(\mathcal{C})}\big(\alpha_1 e^{\mu_1 D}\theta\\ &\qquad-\mu_1 e^{\alpha_1 D}\theta(\tau-2D,\alpha_1)+\lambda P_W(\tau-2D)\big)\\ &\qquad+\frac{(1-\lambda D)e^{-\mu_1(\tau-2D)}}{p(\mathcal{C})}\left(\alpha_1-\mu_1 e^{-\lambda D}+\lambda e^{-\mu_1 D}\right). \end{aligned} \tag{26}$$

D. The Packet Is Not Queued at Either System

Finally, we consider case $\mathcal{D}$, in which there is no queuing at either system. This is equivalent to stating that $T_{i-1,1} \leq Y_{i,1} \wedge \Omega_{i-1,2} + D \leq S_{i,1} - \Omega_{i,1}$. Consequently, we get the following probability for case $\mathcal{D}$:

$$\begin{aligned} p(\mathcal{D}) &= p(\Omega_1 \leq 0, \Omega_2 \leq 0)\\ &= p(Y_i \geq T_{i-1,1}, S_{i,1} \geq T_{i-1,1} - Y_i + W_{i-1,2} + D)\\ &= \int\limits_0^\infty \int\limits_0^{y_1} p_{y_1}(y_1)p_{T_1}(t_1)\\ &\qquad \int\limits_{\max(y_1-t_1-D,0)}^{\infty} (1-P_{S_1}(w+t_1+D-y_1))\\ &\qquad \times p_W(w)dwdt_1\ dy_1 = (1-\lambda D) - p(\mathcal{B}). \end{aligned} \tag{27}$$

The PAoI in this case is given by $Y_{i,1} + S_{i,1} + D$. Since we know that we are in case $\mathcal{D}$, we can apply Bayes' theorem to get:

$$\begin{aligned} &p_{\Delta|\mathcal{D},Y_{i,1},T_{i-1,1}}(\tau|y_1,t_1)\\ &\quad= \frac{p_{S_1}(\tau-y_1-D)P_{Y_1}(y_1)}{p(\mathcal{D})}\\ &\qquad \times P_W(\tau-t_1-2D)u(y_1-t_1)\\ &\quad= \frac{\mu_1\ e^{-\mu_1(\tau-y_1-D)}P_W(\tau-t_1-2D)u(y_1-t_1)}{p(\mathcal{D})}. \end{aligned} \tag{28}$$

We can now uncondition this probability on $Y_{i,1}$ by applying the law of total probability:

$$\begin{aligned} &p_{\Delta|\mathcal{D},T_{i-1,1}}(\tau|t_1)\\ &= \int_t^{\tau-D} \frac{\mu_1\ e^{-\mu_1(\tau-y_1-D)}\lambda e^{-\lambda y_1}}{p(\mathcal{D})}\\ &\quad \times P_W(\tau-t_1-2D)u(y_1-t_1)dy_1\\ &= \frac{\lambda\mu_1}{\alpha_1\ p(\mathcal{D})}P_W(\tau-t_1-2D)e^{-\mu_1(\tau-D)}(e^{\alpha_1(\tau-D)}-e^{\alpha_1 t_1}). \end{aligned} \tag{29}$$

We can now uncondition again on $T_{i-1,i}$ and get the PDF of the PAoI in case $\mathcal{D}$:

$$\begin{aligned} p_{\Delta|\mathcal{D}}(\tau) &= \int_D^{\tau-2D} \frac{\lambda\mu_1}{p(\mathcal{D})}P_W(\tau-t_1-2D)e^{-\mu_1(\tau-D)}\\ &\qquad \times(e^{\alpha_1(\tau-D-t_1)}-1)dt_1\\ &= \frac{\lambda\mu_1}{p(\mathcal{D})}e^{-\mu_1(\tau-D)}\left(e^{\alpha_1\ D}\int_0^{\tau-2D} P_W(w)e^{\alpha_1\ w}dw\right.\\ &\qquad \left.-\int_0^{\tau-2D} P_W(w)dw\right)\\ &= \frac{\mu_1\lambda(1-\lambda D)}{p(\mathcal{D})}e^{-\mu_1(\tau-D)}\sum_{k=0}^{\left\lfloor\frac{\tau-2D}{D}\right\rfloor}\left[\frac{1}{\lambda}-e^{\alpha_1(k+1)D}\right.\\ &\qquad +\sum_{j=0}^{k}\left(\frac{(\tau-(k+2)D)^j}{j!}\left((-\lambda)^{j-1}e^{\lambda(\tau-(k+2)D)}\right.\right.\\ &\qquad \left.\left.\left.-\frac{\lambda^k(-1)^j e^{\mu_1(\tau-(k+2)D)}}{\mu_1^{k-j+1}}\right)\right)\right]. \end{aligned} \tag{30}$$

The overall PAoI PDF is then given by using the previous results for the four cases in Definition 1.

V. PAoI Distribution for the $M/M/1$ – $M/M/1$ Tandem

We now consider the $M/M/1$ – $M/M/1$ tandem, which represents edge computing-enabled systems with stochastic computation times or communication relaying systems. To calculate (11), we divide the computation in 4 cases, as we did in the previous section.

A. The Packet Is Queued at Both Systems

We first consider case $\mathcal{A}$, in which packet i finds both systems busy, i.e., the i-th packet arrives before the departure of the $(i-1)$-th packet at each system. In this case, $\Omega_{i,1} > 0 \wedge \Omega_{i,2} > 0$. As the conditioned PDF of $\Omega_{i,j}$ was given in (10), and we know that $Y_{i,1}$ is independent from $T_{i-1,1}$, as is $S_{i,1}$ from $T_{i-1,2}$, the probability of this case $p(\mathcal{A})$ is given by:

$$\begin{aligned} &p(\mathcal{A})\\ &= p(\Omega_1 > 0)p(\Omega_2 > 0|\Omega_1 > 0)\\ &= \int\limits_0^\infty p_{T_1}(t_1)\int\limits_0^{t_1} p_{y_1}(y_1)dy_1dt_1 \int\limits_0^\infty p_{T_2}(t_2)\int\limits_0^{t_2} p_{S_1}(s_1)ds_1dt_2\\ &= \frac{\lambda}{\mu_1+\alpha_2}. \end{aligned} \tag{31}$$

We start from the conditioned distribution of the system time on Ω_1, Ω_2, and S_1, so S_2 is the only remaining random variable. In the following, the index i of the packet is omitted where possible to simplify the notation:

$$\begin{aligned} p_{T|\Omega_1,\Omega_2,S_1,\mathcal{A}}(t|\omega_1,\omega_2,s_1) &= \mu_2 e^{-\mu_2(t-\omega_1-s_1-\omega_2)}\\ &\qquad \times u(t-\omega_1-\omega_2-s_1). \end{aligned} \tag{32}$$

We now uncondition on Ω_2, and then on S_1, by using the law of total probability:

$$\begin{aligned}&p_{T|\Omega_1,\mathcal{A}}(t|\omega_1)\\&=\int\limits_0^{t-\omega_1} p_{S_1}(s_1)\int\limits_0^{t-s_1-\omega_1}\frac{p_{\Omega_2|\Omega_1,S_1}(\omega_2|\omega_1,s_1)}{1-P_{\Omega_2|\Omega_1,S_1}(0|\omega_1,s_1)}\\&\quad\times p_{T|\Omega_1,\Omega_2,S_1,\mathcal{A}}d\omega_2 ds_1\\&=\frac{\alpha_2\mu_2(\alpha_2+\mu_1)e^{-\alpha_2(t-\omega_1)}(\alpha_1+\lambda e^{-\mu_1(t-\omega_1)}-\mu_1 e^{-\lambda(t-\omega_1)})}{\lambda\alpha_1\mu_1}.\end{aligned}\tag{33}$$

The knowledge that we are in case $\mathcal{A}$ means that we have $\Omega_1 > 0$: the denominator in the first integral is the probability of this happening, which we need to account for to get the correct conditional probability. We then condition on Y_1 and uncondition on Ω_1:

$$\begin{aligned}&p_{T|Y_1,\mathcal{A}}(t|y_1)\\&\quad=\int\limits_0^{t} p_{T|\Omega_1,\mathcal{A}}(t|\omega_1)\frac{p_{\Omega_1|Y_1}(\omega_1|y_1)}{1-P_{\Omega_1|Y_1}(0|y_1)}d\omega_1\\&\quad=\frac{\mu_2\alpha_2\, e^{-\alpha_1\, y_1}}{\lambda p(\mathcal{A})}\Bigg(\frac{\alpha_1(e^{-\alpha_1 t}-e^{-\alpha_2 t})}{(\mu_2-\mu_1)}\\&\qquad+\frac{\lambda e^{-\alpha_1 t}(1-e^{-\mu_2 t})}{\mu_2}-\frac{\mu_1(e^{-\alpha_1 t}-e^{-\mu_2 t})}{\mu_2-\alpha_1}\Bigg).\end{aligned}\tag{34}$$

We can now derive the PDF of the system time T:

$$\begin{aligned}&p_{T|\mathcal{A}}(t)\\&\quad=\int_0^{\infty} p_{Y_1}(y_1)p_{T|Y_1,\mathcal{A}}(t|y_1)dy_1\\&\quad=\frac{\mu_2\alpha_2}{\mu_1\; p(\mathcal{A})}\Bigg(\frac{\alpha_1(e^{-\alpha_1 t}-e^{-\alpha_2 t})}{(\mu_2-\mu_1)}\\&\qquad+\frac{\lambda e^{-\alpha_1 t}(1-e^{-\mu_2 t})}{\mu_2}-\frac{\mu_1(e^{-\alpha_1 t}-e^{-\mu_2 t})}{\mu_2-\alpha_1}\Bigg).\end{aligned}\tag{35}$$

Finally, we get the PDF of the PAoI, given by $T+Y_1$:

$$\begin{aligned}p_{\Delta|\mathcal{A}}(\tau)&=\int_0^{\tau} p_{T|Y_1,\mathcal{A}}(t|\tau-t)\,p_{Y_1}(\tau-t)dt\\&=\frac{\mu_1+\alpha_2}{\lambda}\Bigg(\frac{\alpha_2\mu_1\mu_2(e^{-\mu_1\tau}-e^{-\mu_2\tau})}{(\mu_2-\mu_1)(\mu_2-\alpha_1)}\\&\quad-\lambda e^{-\mu_1\tau}(1-e^{-\alpha_2\tau})\\&\quad+\frac{\alpha_1\alpha_2\mu_2(e^{-\mu_1\tau}-e^{-\alpha_2\tau})}{(\mu_2-\mu_1)(\mu_1-\alpha_2)}\\&\quad+\frac{\alpha_1\mu_1\mu_2(e^{-\alpha_1\tau}-e^{-\mu_1\tau})}{(\mu_2-\mu_1)(\mu_2-\alpha_1)}\Bigg).\end{aligned}\tag{36}$$

B. The Packet Is Only Queued at the First System

We now consider case $\mathcal{B}$, in which the first system is busy but the second one is free when packet i reaches it, i.e., the packet is not queued at the second system. We have $\Omega_{i,1} > 0 \wedge \Omega_{i,2} \leq 0$, and this case happens with probability $p(\mathcal{B})$:

$$\begin{aligned}&p(\mathcal{B})\\&=p(\Omega_1>0)p(\Omega_2\leq 0|\Omega_1>0)\\&=\int\limits_0^{\infty}\int\limits_0^{t_1} p_{y_1}(y_1)p_{T_1}(t_1)dy_1dt_1\int\limits_0^{\infty}\int\limits_{t_2}^{\infty} p_{S_1}(s_1)p_{T_2}(t_2)ds_1dt_2\\&=\frac{\lambda\alpha_2}{\mu_1(\mu_1+\alpha_2)}.\end{aligned}\tag{37}$$

In this case, the system time PDF is independent of Ω_2, and we can just give the conditioned PDF as:

$$\begin{aligned}p_{T|\Omega_1,S_1,\mathcal{B}}(t|\omega_1,s_1)=\mu_2 e^{-\mu_2(t-\omega_1-s_1)}(1-e^{-\alpha_2 s_1})\\\times u(t-\omega_1-s_1).\end{aligned}\tag{38}$$

As for case $\mathcal{A}$, we condition on Y_1 and uncondition on S_1 and Ω_1:

$$\begin{aligned}&p_{T|Y_1,\mathcal{B}}(t|y_1)\\&\qquad=\frac{\mu_1(\mu_1+\alpha_2)e^{-\alpha_1(y_1+t)}}{\alpha_2}\Bigg(1-e^{-\mu_2 t}\\&\qquad\quad-\frac{\alpha_2\mu_2(1-e^{(\alpha_1-\mu_2)t})}{(\mu_2-\mu_1)(\mu_2-\alpha_1)}+\frac{\alpha_1\mu_2(1-e^{-\lambda t})}{\lambda(\mu_2-\mu_1)}\Bigg).\end{aligned}\tag{39}$$

From this result, we derive the conditioned PDF of the system time T for case $\mathcal{B}$:

$$\begin{aligned}&p_{T|\mathcal{B}}(t)\\&\quad=\frac{\lambda e^{-\alpha_1 t}}{p(\mathcal{B})}\Bigg(1-e^{-\mu_2 t}\\&\qquad+\frac{\alpha_1\mu_2(1-e^{-\lambda t})}{\lambda(\mu_2-\mu_1)}-\frac{\alpha_2\mu_2\left(1-e^{-(\mu_2-\alpha_1)t}\right)}{(\mu_2-\mu_1)(\mu_2-\alpha_1)}\Bigg).\end{aligned}\tag{40}$$

We can now find the unconditioned PDF of the PAoI for case $\mathcal{B}$ (41), as shown at the bottom of the next page.

C. The Packet Is Only Queued at the Second System

We can then consider case $\mathcal{C}$, in which the i-th packet does not experience any queuing at the first system, i.e., $\Omega_1 \leq 0$, but there is queuing in the second system, i.e., $\Omega_2 > 0$. The probability of a packet experiencing case $\mathcal{C}$ is given by:

$$\begin{aligned}p(\mathcal{C})&=p(\Omega_1\leq 0,\Omega_2>0)\\&=\int\limits_0^{\infty}\int\limits_0^{y_1}\int\limits_0^{\infty} p_{y_1}(y_1)p_{T_1}(t_1)p_{S_1}(s_1)\\&\qquad\int\limits_{s_1-t_1+y_1}^{\infty} p_{T_2}(t_2)dt_2ds_1dt_1dy_1\\&=\frac{\lambda}{\mu_2(\mu_1+\alpha_2)}.\end{aligned}\tag{42}$$

The conditioned PDF of the system time is:

$$p_{T|\Omega_1,\Omega_2,S_1,\mathcal{C}}(t|\omega_1,\omega_2,s_1)=\mu_2 e^{-\mu_2(t-s_1-\omega_2)}u(t-s_1-\omega_2).\tag{43}$$

As in case $\mathcal{A}$, we condition on Y_1 and uncondition on Ω_2, S_1, and Ω_1:

$$p_{T|Y_1,\mathcal{C}}\left(t|y_1\right) = \frac{\mu_2\alpha_2 e^{-\alpha_2 t}(e^{-\alpha_1 y_1} - e^{-\alpha_2 y_1})}{\lambda(\mu_2-\mu_1)p(\mathcal{C})} \times \left(\alpha_1 - \mu_1 e^{-\lambda t} + \lambda e^{-\mu_1 t}\right). \quad (44)$$

We can now find the PDF of the system delay:

$$p_{T|\mathcal{C}}(t) = \frac{\alpha_2 e^{-\alpha_2 t}\left(\alpha_1 - \mu_1 e^{-\lambda t} + \lambda e^{-\mu_1 t}\right)}{\mu_1\ p(\mathcal{C})}. \quad (45)$$

The conditioned PDF of the PAoI is then:

$$\begin{aligned} &p_{\Delta|\mathcal{C}}(\tau) \\ &= \frac{\mu_2}{(\mu_2-\mu_1)p(\mathcal{C})}\left(\frac{\alpha_1\alpha_2(e^{-\mu_1\tau} - e^{-\alpha_2\tau})}{\alpha_2-\mu_1}\right. \\ &\quad +\lambda e^{-\mu_1\tau} - \frac{\mu_1\alpha_2(e^{-\mu_1\tau} - e^{-\mu_2\tau})}{\mu_2-\mu_1} \\ &\quad -\frac{\alpha_1\alpha_2(e^{-\alpha_2\tau} - e^{-\mu_2\tau})}{\lambda} \\ &\quad \left.-\lambda e^{-(\mu_1+\alpha_2)\tau} + \alpha_2\mu_1\tau e^{-\mu_2\tau} - \frac{\lambda\alpha_2 e^{-\mu_2\tau}(1-e^{-\alpha_1\tau})}{\alpha_1}\right). \end{aligned} \quad (46)$$

D. The Packet Is Not Queued at Either System

Finally, we examine case $\mathcal{D}$, in which the packet experiences no queuing, i.e., $\Omega_{i,1} \leq 0 \wedge \Omega_{i,2} \leq 0$. This case happens with probability $p(\mathcal{D})$:

$$\begin{aligned} p(\mathcal{D}) &= p(\Omega_1 \leq 0, \Omega_2 \leq 0) \\ &= \frac{\alpha_1\mu_2(\mu_1+\alpha_2)) - \lambda\mu_1}{\mu_1\mu_2(\mu_1+\alpha_2)}. \end{aligned} \quad (47)$$

Since the system time probability is independent of Ω_2, we can just give the conditioned system time PDF as:

$$p_{T|\Omega_1,S_1,\mathcal{D}}\left(t|\omega_1,s_1\right) = \frac{\mu_1\mu_2 e^{-\mu_2(t-s_1)}(1-e^{-\alpha_2(s_1-\omega_1)})}{\alpha_1} \times u(t-s_1). \quad (48)$$

We then condition on Y_1 and uncondition on S_1 and Ω_1:

$$\begin{aligned} &p_{T|Y_1,\mathcal{D}}\left(t|y_1\right) \\ &= \frac{\mu_1\mu_2}{(\mu_2-\mu_1)\big(p(\mathcal{D})} e^{-\mu_1 t}(1-e^{-\alpha_1 y_1}) \\ &\quad -e^{-\mu_2 t}(1-e^{-\alpha_2 y_1}) + e^{-(\mu_2+\alpha_1)t}(e^{-\alpha_1 y_1} - e^{-\alpha_2 y_1})\big). \end{aligned} \quad (49)$$

The PDF of the system time is:

$$p_{T|\mathcal{D}}(t) = \frac{\mu_2(\mu_1-\lambda)e^{-\mu_1 t} - \mu_1(\mu_2-\lambda)e^{-\mu_2 t}}{\lambda(\mu_2-\mu_1)p(\mathcal{D})} + \frac{\lambda e^{-(\mu_2+\alpha_1)t}}{p(\mathcal{D})}. \quad (50)$$

We can now find the PDF of the PAoI in case $\mathcal{D}$:

$$\begin{aligned} &p_{\Delta|\mathcal{D}}(\tau) \\ &= \frac{\mu_1\mu_2\lambda}{p(\mathcal{D})}\left(\frac{\tau(e^{-\mu_2\tau} - e^{-\mu_1\tau})}{\mu_2-\mu_1}\right. \\ &\quad \left.+\frac{\alpha_1 e^{-\mu_2\tau} - \alpha_2 e^{-\mu_1\tau}}{\alpha_1\alpha_2(\mu_2-\mu_1)} + \frac{\left(e^{-\lambda\tau} + e^{-(\mu_1+\alpha_2)\tau}\right)}{\alpha_1\alpha_2}\right). \end{aligned} \quad (51)$$

As for the $M/M/1$ – $M/D/1$ system, the total PDF is given by Theorem 1.

E. PAoI in the Case with Equal Service Rates

In this subsection, we consider a special case in which the general formula of the PAoI PDF is indeterminate, $\mu_1 = \mu_2 = \mu$. We follow the same steps as in the normal derivation. The other cases in which the general formula is indeterminate, $\mu_1 = \alpha_2$ and $\mu_2 = \alpha_1$, are not derived in this paper.

In case $\mathcal{A}$, i.e., for $\Omega_1 > 0 \wedge \Omega_2 > 0$, we have:

$$p(\mathcal{A}) = \frac{\lambda}{\mu+\alpha}. \quad (52)$$

Following the same steps as in the general case, we get:

$$\begin{aligned} p_{\Delta|\mathcal{A}}\left(\tau\right) = \frac{e^{-\mu\tau}}{p(\mathcal{A})}&\left(\mu(e^{\lambda\tau}-1) + \lambda(e^{-\alpha\tau} - e^{\lambda\tau})\right. \\ &\left.+\frac{\mu\alpha(\alpha+\mu)(1-e^{\lambda\tau}) + \mu\alpha\lambda\tau(\alpha e^{\lambda\tau}+\mu)}{\lambda^2}\right). \end{aligned} \quad (53)$$

The probability of being in case $\mathcal{B}$, i.e., $\Omega_1 > 0 \wedge \Omega_2 \leq 0$, is:

$$p(\mathcal{B}) = \frac{\lambda\alpha}{\mu(\mu+\alpha)}. \quad (54)$$

The conditioned PDF of the PAoI is then:

$$\begin{aligned} p_{\Delta|\mathcal{B}}\left(\tau\right) = \frac{\mu e^{-\mu\tau}}{p(\mathcal{B})}&\left(\frac{\alpha^2(e^{\lambda\tau}-1)}{\lambda^2} - \frac{\lambda(1-e^{-\alpha\tau})}{\alpha}\right. \\ &\left.-\frac{\mu(\alpha\lambda\tau^2 + (\alpha-\lambda)\tau)}{2\lambda}\right). \end{aligned} \quad (55)$$

$$\begin{aligned} p_{\Delta|\mathcal{B}}\left(\tau\right) = &\frac{\mu_1}{p(\mathcal{B})}(e^{-\alpha_1\tau} - e^{-\mu_1\tau}) - \frac{\lambda\mu_1 e^{-\mu_1\tau}(1-e^{-\alpha_2\tau})}{\alpha_2 p(\mathcal{B})} \\ &+\frac{\alpha_1\mu_1\mu_2(e^{-\alpha_1\tau} - (1+\lambda\tau)e^{-\mu_1\tau})}{\lambda(\mu_2-\mu_1)p(\mathcal{B})} \\ &+\frac{\mu_1\mu_2\alpha_2\left((e^{-\mu_1\tau} - e^{-\alpha_1\tau})(\mu_2-\mu_1) + \lambda(e^{-\mu_1\tau} - e^{-\mu_2\tau})\right)}{(\mu_2-\mu_1)^2(\mu_2-\alpha_1)p(\mathcal{B})}. \end{aligned} \quad (41)$$

TABLE II
MAIN SIMULATION PARAMETER VALUES (UNLESS OTHERWISE SPECIFIED)

Parameter	Value	Description
λ	0.5	Packet generation rate
μ_1	1	Service rate of $M/M/1$ system 1
μ_2	1.25	Service rate of $M/M/1$ system 2
D	0.8	Service time of the $M/D/1$ system
N	10^7	Number of simulated packets
N_0	1000	Initial transition

In this case, as the system is entirely symmetrical and both queues are $M/M/1$, it is also time reversible, making case $\mathcal{B}$ equivalent to case $\mathcal{C}$ in reverse. The probability of being in case $\mathcal{C}$, i.e., $\Omega_1 \leq 0 \wedge \Omega_2 > 0$, and the conditional PDF of the PAoI are then the same as in case $\mathcal{B}$:

$$p(\mathcal{C}) = p(\mathcal{B}) p_{\Delta|\mathcal{C}}(\tau) = p_{\Delta|\mathcal{B}}(\tau) \,. \qquad (56)$$

Finally, we look at case $\mathcal{D}$, in which both systems are free:

$$p(\mathcal{D}) = \frac{\alpha(\mu+\alpha) - \lambda}{\mu(\mu+\alpha)}. \qquad (57)$$

We then have the conditioned PDF of the PAoI:

$$p_{\Delta|\mathcal{D}}(\tau) = \frac{\mu^2 \lambda e^{-\mu\tau}(2\cosh(\alpha\tau) - \alpha^2\tau^2 - 2)}{\alpha^2 \; p(\mathcal{D})}. \qquad (58)$$

As in the general case, the overall PAoI is given by Theorem 1.

VI. SIMULATION RESULTS

We compared the results of our analysis with a Monte Carlo simulation, transmitting $N = 10^7$ packets and computing the system delay and PAoI for each. The initial stages of each simulation were discarded, removing $N_0 = 1000$ packets to ensure that the system had reached a steady state. We also divided the packets in the four cases, depending on the queuing they experienced at each system. As the derivation of the PAoI distribution does not involve any approximations, the simulation results should perfectly match the theoretical curves. The Monte Carlo simulation consisted of a single episode, with all packets being transmitted one after the other. The simulation parameters are listed in Table II, and are used for all plots, unless otherwise specified.

A. $M/M/1$ – $M/D/1$ Tandem

We first consider the tandem in which the first queue is $M/M/1$, while the second is $M/D/1$. We note that, in all the following figures, the simulation results match the theoretically derived curves, showing the soundness of our calculations.

Although not shown, the system time has the expected behavior: it is highest in case $\mathcal{A}$ when the packet is queued at both, and lowest in case $\mathcal{D}$, in which both systems are free. However, the PAoI shows a different trend. While system time increases monotonically with the traffic load, the PAoI is the combination of the system and interarrival time: at one extreme, when the system has very low traffic, it is dominated by the interarrival time, while at the other, it is dominated by the system time. The optimal setting to minimize PAoI is somewhere in the middle, striking a balance between the two causes of age. Furthermore, deterministic service reduces uncertainty, particularly when traffic is high and queuing is the main cause of ageing.

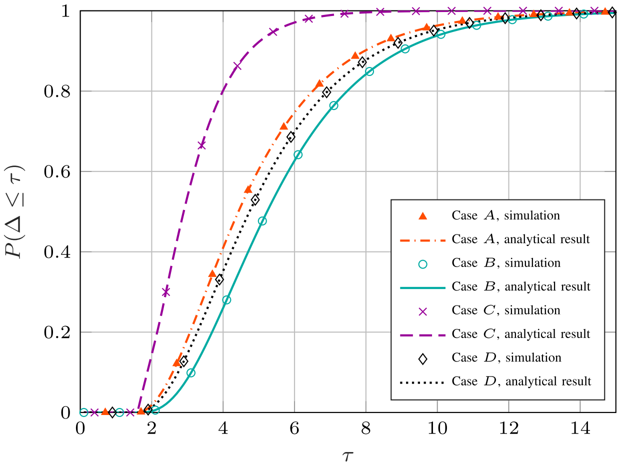


Fig. 5. CDF of the PAoI Δ for the $M/M/1$ – $M/D/1$ tandem in the four subcases.

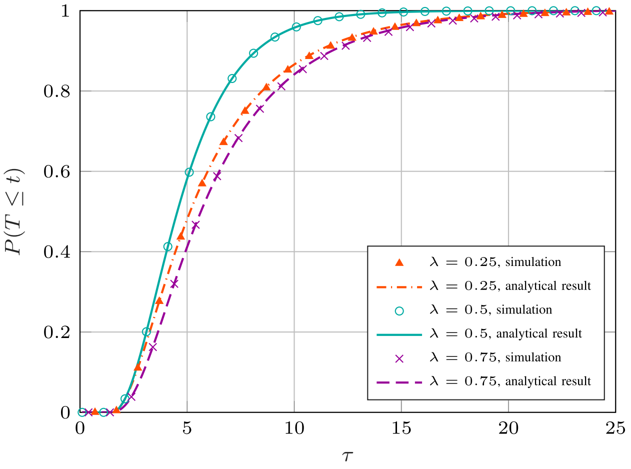


Fig. 6. CDF of the PAoI Δ for the $M/M/1$ – $M/M/1$ tandem for different values of λ.

Fig. 5 shows the CDF of the PAoI in the four cases. It is interesting to note that the PAoI is never smaller than $2D$, as even serving packets instantaneously at the first system would still lead to a minimum delay: once packet $i-1$ is generated, it needs at least a time D to get through the system because of the $M/D/1$ queue, and even if packet i is generated right after it it needs another D to be served by the edge node, leading to a minimum age of $2D$. The PAoI is far smaller in case $\mathcal{C}$, i.e., when the packet is queued only at the $M/D/1$ system, as the queue will often be short and is guaranteed to empty in a limited time. Cases $\mathcal{A}$ and $\mathcal{B}$ show a far worse performance, because of the first system's lower service rate ($D = 0.8$ corresponds to a rate of 1.25) and of its exponential system time distribution. If the packet is not queued at either system (case $\mathcal{D}$), the PAoI is dominated by the interarrival time between the two packets, leading to higher ages.

We can also examine the PAoI as a function of the generation and service rates: Fig. 6 shows the PAoI CDF for different values of λ. We can observe that a value of λ close

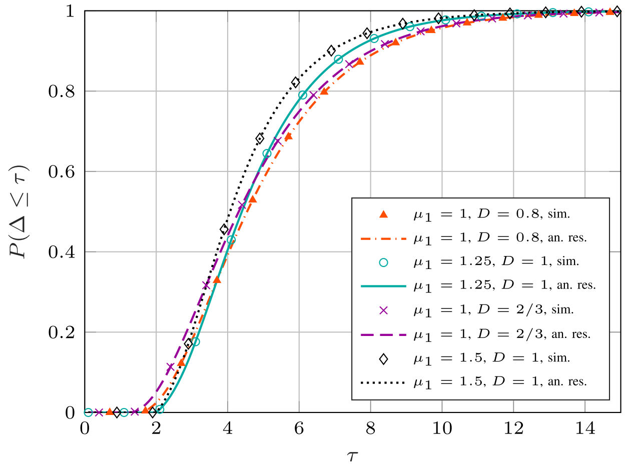


Fig. 7. CDF of the PAoI Δ for the $M/M/1$ – $M/M/1$ tandem for different values of μ_1 and D.

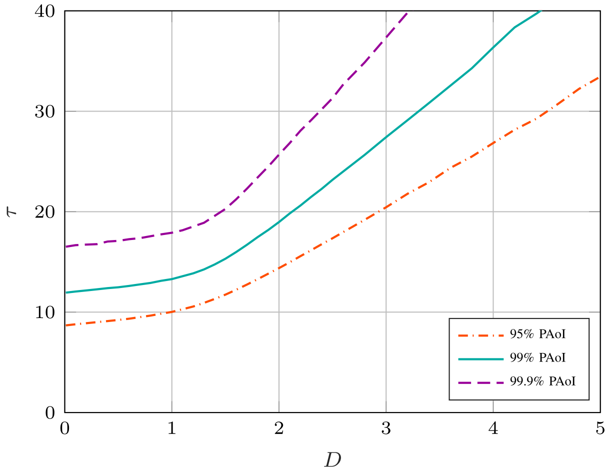


Fig. 9. Tail of the CDF for the $M/M/1$ – $M/D/1$ tandem with $\mu_1 = 1$, for different values of D using the optimal λ.

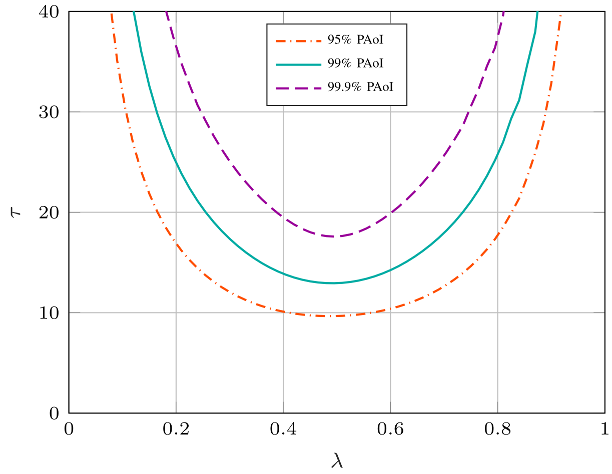


Fig. 8. Tail of the CDF for the $M/M/1$ – $M/D/1$ tandem with $\mu_1 = 1$ and $D = 0.8$, for different values of λ.

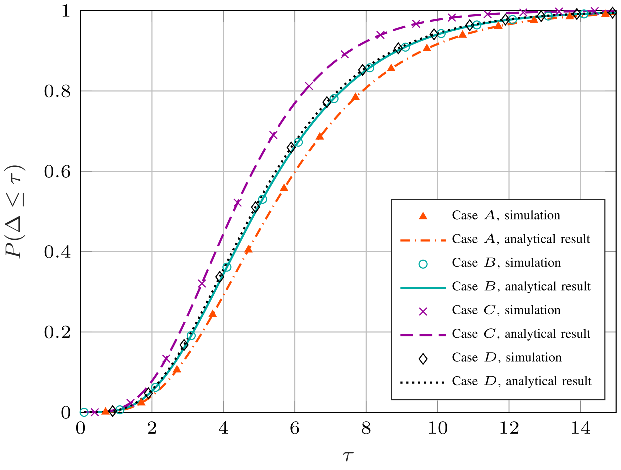


Fig. 10. CDF of the PAoI Δ for the $M/M/1$ – $M/M/1$ tandem in the four subcases.

to 0.5 leads to the lowest PAoI, as a high traffic load increases queuing times, while a lower load increases PAoI due to the longer interarrival times. Fig. 7 shows what happens when the service rates of the two systems are flipped.

In the figure, we compare two pairs of curves (orange and cyan, violet and black): the orange and violet dashed lines correspond to systems where the second node is faster, the cyan solid line and the black dotted line have the $M/D/1$ queue as the bottleneck. As the system time in the $M/D/1$ queues is rarely very large (it would need a very large queue to be significant, as all packets have the same service time), while $M/M/1$ queues have an exponential system time distribution which can take larger values more often, placing the bottleneck on the $M/D/1$ system leads to a lower PAoI in the worst case (i.e., the larger percentiles), at the cost of a worse PAoI in favorable scenarios. The difference between the orange and black lines is larger than between the cyan and violet ones, as a larger difference between the rates of the two links leads to an increased importance of the bottleneck. As for the subcase analysis, the system time and PAoI from the Monte Carlo simulation follow the analytical curve perfectly. As Fig. 9 shows, improving the edge computing capabilities has diminishing returns, as the $M/M/1$ communication system becomes the bottleneck: even with a very low D, the PAoI cannot be reduced beyond a certain value without also improving the first system's capacity.

We also make a worst-case analysis as a function of λ: Fig. 8 shows the 95th, 99th and 99.9th percentiles of the PAoI for a system with $\mu_1 = 1$ and $D = 0.8$. If the traffic is very high, the queuing time is the dominant factor, causing the worst-case PAoI to diverge. The same happens if the traffic is too low, as the interarrival times can be very large: in this case, the system will almost always be empty, but updates will be very rare. The best performance in terms of PAoI is close to the middle. Depending on the desired reliability, system designers should choose the range of λ that fulfils the given percentile of the PAoI in the specific conditions they consider.

B. $M/M/1$ – $m/M/1$ Tandem

Fig. 10 shows the PAoI CDF in the four subcases for $\lambda = 0.5$, $\mu_1 = 1$, and $\mu_2 = 1.2$. Interestingly, there is no minimum delay and the CDF starts in zero, unlike in the

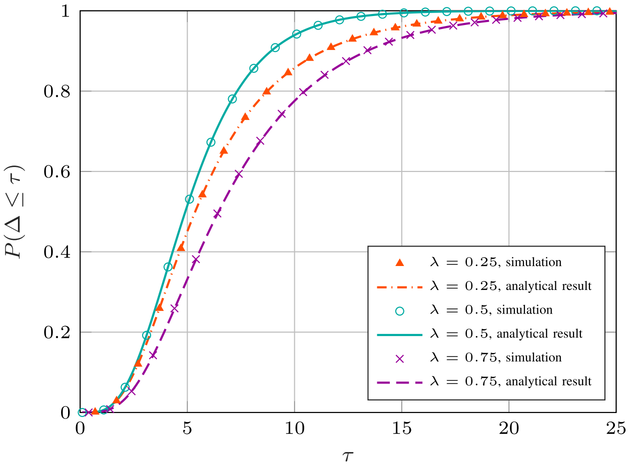


Fig. 11. CDF of the PAoI Δ for the $M/M/1 - M/M/1$ tandem for different values of λ.

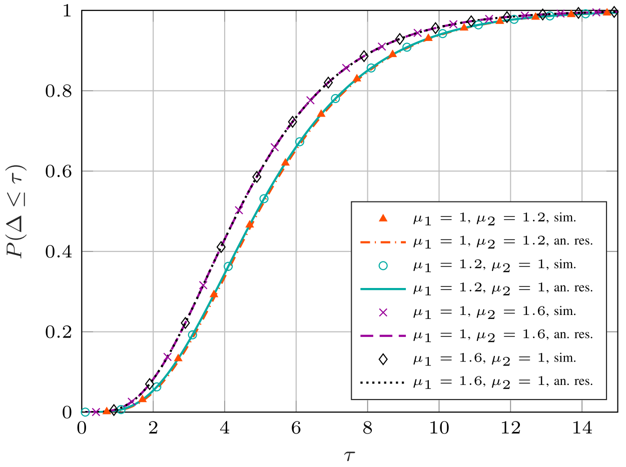


Fig. 12. CDF of the PAoI Δ for the $M/M/1 - M/M/1$ tandem for different values of μ_1 and μ_2.

$M/M/1$–$M/D/1$ case. The PAoI is the lowest in case $\mathcal{C}$, and almost identical in cases B and D. This is due to the effect of the interarrival times on the PAoI, as case $\mathcal{D}$ usually means that the instantaneous load of the system is low and packets are far apart, increasing the PAoI. In case $\mathcal{C}$, the faster system is busy and the bottleneck is empty. Intuitively, this can reduce age, as the second system will probably be able to serve packets fast enough, but at the same time the instantaneous load will be high enough to avoid having a strong impact on the age.

We can now examine the PAoI CDFs for different values of λ: the system time is always higher for higher values of λ, as it depends on the traffic. The same is not true for the PAoI, as Fig. 11 shows: as for the $M/M/1 - M/D/1$ tandem, the PAoI is lowest for $\lambda = 0.5$, as the high interarrival time becomes the dominant factor for $\lambda = 0.25$. Interestingly, the gap between the system with $\lambda = 0.25$ and the one with $\lambda = 0.75$ is wider: the $M/D/1$ system is better able to handle high load situations, as having a high system time is far rarer than in the $M/M/1$. On the other hand, the values of μ_1 and μ_2 also have an important effect, as Fig. 12 shows: while the bottleneck always has a service rate 1, changing the service rate of the other link and even switching the two can have an impact on the PAoI. Naturally, increasing the rate of the other link from 1.2 to 1.6 slightly reduces the PAoI, but we note that for both values, having the first or second system as the bottleneck has a negligible effect on performance. Unlike in the $M/M/1 - M/D/1$ tandem, the location of the bottleneck in the tandem seems to have a very small influence on the distribution of the PAoI, as both queues have the same type of service time distribution.

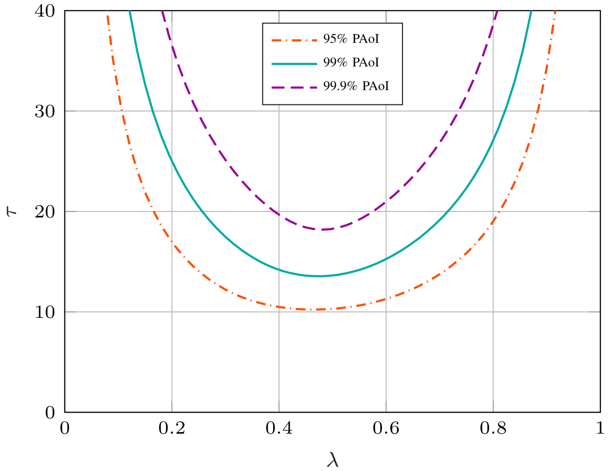


Fig. 13. Tail of the CDF for the $M/M/1 - M/M/1$ tandem with $\mu_1 = 1$ and $\mu_2 = 1.25$, for different values of λ.

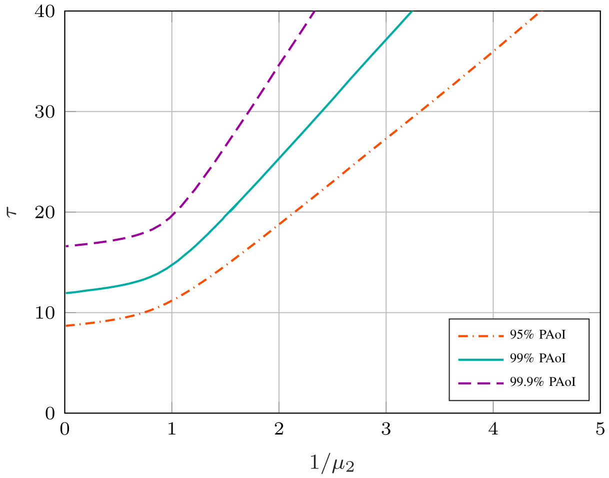


Fig. 14. Tail of the CDF for the $M/M/1 - M/M/1$ tandem with $\mu_1 = 1$, for different values of μ_2 and using the optimal λ.

Fig. 13 shows how the worst-case PAoI, measured using the 95th, 99th and 99.9th percentiles, changes as a function of λ. The figure shows that the trends for the two systems are similar: the $M/M/1 - M/D/1$ queue can handle a high load slightly better, and its minimum PAoI is lower by 5-10% at all the considered percentiles. In both cases, the optimal λ is between 0.45 and 0.5 for all the three percentiles. As for the $M/M/1 - M/D/1$ tandem, increasing the rate of the second system gives diminishing returns, and any increase past $\mu_2 = 2$ has negligible benefits if $\mu_1 = 1$, as Fig. 14 shows. The horizontal axis in the figure shows the inverse of μ_2 to provide a visual comparison with Fig. 8: the tail of the PAoI distribution clearly has higher values in the system with two

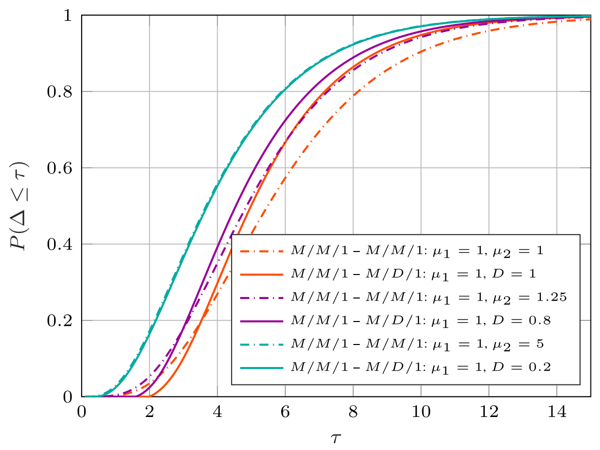


Fig. 15. CDF of the PAoI Δ for the two systems (analytical results) for different values of μ_1 and D.

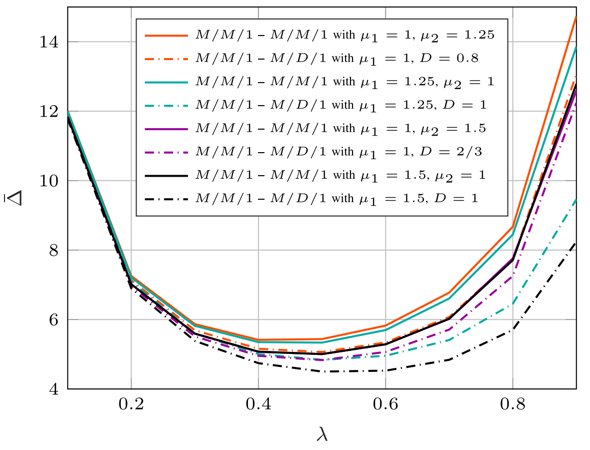


Fig. 16. Average PAoI $\bar{\Delta}$ for the two systems (analytical results) for different values of μ_1 and D.

$M/M/1$ queues, as we would expect due to the additional randomness in the service time of the edge computing node.

Finally, Fig. 15 shows a comparison of the two kinds of system: as expected, the $M/M/1$ – $M/D/1$ tandem has a higher PAoI in the best-case scenario, as it has a hard minimum of $2D$, but quickly becomes better at the higher percentiles. This difference is more pronounced if the second system is slower, with the $M/M/1$ – $M/D/1$ tandem with $D = 1$ having a better PAoI than the $M/M/1$ – $M/M/1$ one with $\mu_2 = 1.25$ for percentiles above the 60th. As the plot shows, the two types of tandem are substantially equivalent if the computation is much faster than the communication. We also show the average PAoI for the two systems in Fig. 16: this is computable from the PDF we derived in this paper, but simpler formulas were derived in the literature [23], [28]. As expected, the $M/M/1$–$M/D/1$ system is better able to deal with high traffic load than the $M/M/1$–$M/M/1$. It is interesting to note that the $M/M/1$–$M/D/1$ queue has a larger gain in the average PAoI than at higher percentiles, even for values of λ between 0.45 and 0.5.

VII. Conclusion and Future Work

In this paper, we derived the PDF of the PAoI for a tandem with an $M/M/1$ system followed by an $M/D/1$ and for one consisting of two $M/M/1$ queues. These new results can give more flexibility in the design of bounded AoI systems, both for edge-enabled IoT where the wireless tranmission precedes the computation delay and in other relay applications. The results are derived for two nodes, but the procedure is generic for K $M/M/1$ nodes, potentially followed by an $M/D/1$ system. If the $M/D/1$ system is not the last, the analytical derivation of the PAoI becomes intractable, as its departure process is non-Markovian.

The optimization of the two systems can be a complex operation to perform in real time, and finding the complete PDF analytically might not be manageable in more complex cases with channel errors and multiple sources. However, approximations of the optimal policies can be found by applying deep learning techniques [51], paying particular attention to the uncertainty in the model parameters in order to guarantee reliability even in realistic conditions [52]. The possibility to compute the PDF and CDF using the formulas can provide a fast way to generate samples for a deep learning system, which would then be able to generalize experience in previously unseen situations with a lower computational cost: there is an extensive body of work on similar approaches, which use learning to quickly find a solution to optimization problems with computationally expensive cost functions.

Aside from the possibilities offered by deep learning, a possible avenue of future work is the introduction of multiple independent sources in the system, possibly with different priorities. The extension of the system to longer or more complex queuing networks is also a possibility, but the complexity of the derivation might make the analytical results unwieldy. The inclusion of error-prone links, with packets being randomly dropped, is another interesting extension. Finally, the optimization of λ in all these scenarios might be a practical application of the theoretical derivations, yielding usable scheduling policies.

References

[1] M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, “On the role of age of information in the Internet of Things,” *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 72–77, Dec. 2019.

[2] S. Kaul, R. Yates, and M. Gruteser, “Real-time status: How often should one update?” in *Proc. Int. Conf. Comput. Commun.*, Mar. 2012, pp. 2731–2735.

[3] W. Yu *et al.*, “A survey on the edge computing for the Internet of Things,” *IEEE Access*, vol. 6, pp. 6900–6919, 2018.

[4] W. Saad, M. Bennis, and M. Chen, “A vision of 6G wireless systems: Applications, trends, technologies, and open research problems,” *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[5] Q. Kuang, J. Gong, X. Chen, and X. Ma, “Age-of-information for computation-intensive messages in mobile edge computing,” in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019, pp. 1–6.

[6] C. Li, S. Li, and Y. T. Hou, “A general model for minimizing age of information at network edge,” in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2019, pp. 118–126.

[7] N. Abramson, “The ALOHA system: Another alternative for computer communications,” in *Proc. Fall Joint Comput. Conf.*, Nov. 1970, pp. 281–285.

[8] J. Goseling, Č. Stefanović, and P. Popovski, “A pseudo-Bayesian approach to sign-compute-resolve slotted ALOHA,” in *Proc. Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2092–2096.

[9] H. Mroue, A. Nasser, S. Hamrioui, B. Parrein, E. Motta-Cruz, and G. Rouyer, “MAC layer-based evaluation of IoT technologies: LoRa, SigFox and NB-IoT,” in *Proc. IEEE Middle East North Afr. Commun. Conf. (MENACOMM)*, Apr. 2018, pp. 1–5.

[10] X. Song, X. Qin, Y. Tao, B. Liu, and P. Zhang, "Age based task scheduling and computation offloading in mobile-edge computing systems," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshop (WCNCW)*, Apr. 2019, pp. 1–6.

[11] L. Green, "A queueing system in which customers require a random number of servers," *Oper. Res.*, vol. 28, no. 6, pp. 1335–1346, Dec. 1980.

[12] M. Toyoshima, T. Jono, K. Nakagawa, and A. Yamamoto, "Optimum intersatellite link design in the presence of random pointing jitter for free-space laser communication systems," *Proc. SPIE*, vol. 4635, pp. 95–102, Apr. 2002.

[13] S. Lee, M. Kim, J. Lee, R.-H. Hsu, and T. Q. S. Quek, "Is blockchain suitable for data freshness?—Age-of-information perspective," 2020, *arXiv:2006.02735*. [Online]. Available: https://arxiv.org/abs/2006.02735

[14] A. Rovira-Sugranes and A. Razi, "Optimizing the age of information for blockchain technology with applications to IoT sensors," *IEEE Commun. Lett.*, vol. 24, no. 1, pp. 183–187, Jan. 2020.

[15] N. Hassan, S. Gillani, E. Ahmed, I. Ibrar, and M. Imran, "The role of edge computing in Internet of Things," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 110–115, Nov. 2018.

[16] M. S. Elbamby, M. Bennis, W. Saad, M. Latva-Aho, and C. S. Hong, "Proactive edge computing in fog networks with latency and reliability guarantees," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 209, Dec. 2018.

[17] M. S. Elbamby *et al.*, "Wireless edge computing with latency and reliability guarantees," *Proc. IEEE*, vol. 107, no. 8, pp. 1717–1737, Aug. 2019.

[18] Y. Hu and A. Schmeink, "Delay-constrained communication in edge computing networks," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.

[19] J. Gong, Q. Kuang, and X. Chen, "Joint transmission and computing scheduling for status update with mobile edge computing," Feb. 2020, *arXiv:2002.09719*. [Online]. Available: http://arxiv.org/abs/2002.09719

[20] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 869–904, 2nd Quart., 2020.

[21] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: Deep learning for the Internet of Things with edge computing," *IEEE Netw.*, vol. 32, no. 1, pp. 96–101, Jan. 2018.

[22] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Age-optimal information updates in multihop networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 576–580.

[23] C. Xu, H. H. Yang, X. Wang, and T. Q. S. Quek, "Optimizing information freshness in computing-enabled IoT networks," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 971–985, Feb. 2020.

[24] N. Pappas, J. Gunnarsson, L. Kratz, M. Kountouris, and V. Angelakis, "Age of information of multiple sources with queue management," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 5935–5940.

[25] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Queue management for age sensitive status updates," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 330–334.

[26] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Age of information performance of multiaccess strategies with packet management," *J. Commun. Netw.*, vol. 21, no. 3, pp. 244–255, Jun. 2019.

[27] R. D. Yates, "Age of information in a network of preemptive servers," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 118–123.

[28] C. Kam, J. P. Molnar, and S. Kompella, "Age of information for queues in tandem," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2018, pp. 1–6.

[29] N. Akar, O. Dogan, and E. U. Atay, "Finding the exact distribution of (Peak) age of information for queues of PH/PH/1/1 and M/PH/1/2 type," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5661–5672, Sep. 2020.

[30] O. Ayan, H. M. Gursu, A. Papa, and W. Kellerer, "Probability analysis of age of information in multi-hop networks," *IEEE Netw. Lett.*, vol. 2, no. 2, pp. 76–80, Jun. 2020.

[31] B. Buyukates, A. Soysal, and S. Ulukus, "Age of information in multihop multicast networks," *J. Commun. Netw.*, vol. 21, no. 3, pp. 256–267, Jun. 2019.

[32] J. Li, Y. Zhou, and H. Chen, "Age of information for multicast transmission with fixed and random deadlines in IoT systems," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8178–8191, Sep. 2020.

[33] 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. Sel. Areas Commun.*, vol. 37, no. 4, pp. 721–734, Apr. 2019.

[34] H. B. Beytur, S. Baghaee, and E. Uysal, "Measuring age of information on real-life connections," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4.

[35] A. Maatouk, M. Assaad, and A. Ephremides, "On the age of information in a CSMA environment," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 818–831, Apr. 2020.

[36] R. D. Yates and S. K. Kaul, "Age of information in uncoordinated unslotted updating," Feb. 2020, *arXiv:2002.02026*. [Online]. Available: http://arxiv.org/abs/2002.02026

[37] X. Chen, K. Gatsis, H. Hassani, and S. S. Bidokhti, "Age of information in random access channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1770–1775.

[38] 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," Mar. 2020, *arXiv:2003.10916*. [Online]. Available: http://arxiv.org/abs/2003.10916

[39] B. Zhou and W. Saad, "Joint status sampling and updating for minimizing age of information in the Internet of Things," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7468–7482, Nov. 2019.

[40] 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.

[41] F. Chiariotti, O. Vikhrova, B. Soret, and P. Popovski, "Information freshness of updates sent over LEO satellite multi-hop networks," Jul. 2020, *arXiv:2007.05449*. [Online]. Available: http://arxiv.org/abs/2007.05449

[42] J. P. Champati, H. Al-Zubaidy, and J. Gross, "Statistical guarantee optimization for AoI in single-hop and two-hop systems with periodic arrivals," Oct. 2019, *arXiv:1910.09949*. [Online]. Available: http://arxiv.org/abs/1910.09949

[43] C.-F. Liu and M. Bennis, "Taming the tail of maximal information age in wireless industrial networks," *IEEE Commun. Lett.*, vol. 23, no. 12, pp. 2442–2446, Dec. 2019.

[44] C. Chaccour and W. Saad, "On the ruin of age of information in augmented reality over wireless teraherts (THz) networks," in *Proc. Globecom*, Dec. 2020, pp. 1–6.

[45] B. Zhou, W. Saad, M. Bennis, and P. Popovski, "Risk-aware optimization of age of information in the Internet of Things," Feb. 2020, *arXiv:2002.09805*. [Online]. Available: http://arxiv.org/abs/2002.09805

[46] R. D. Yates, Y. Sun, D. R. Brown, III, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," Jul. 2020, *arXiv:2007.08564*. [Online]. Available: http://arxiv.org/abs/2007.08564

[47] E. Reich, "Note on queues in tandem," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 338–341, Mar. 1963.

[48] P. J. Burke, "The output of a queuing system," *Oper. Res.*, vol. 4, no. 6, pp. 699–704, Dec. 1956.

[49] A. Erlang, "Telefon-ventetider. Et stykke sandsynlighedsregning," *Matematisk tidsskrift. B*, vol. 31, pp. 25–42, Jan. 1920.

[50] V. B. Iversen and L. Staalhagen, "Waiting time distribution in M/D/1 queueing systems," *Electron. Lett.*, vol. 35, no. 25, pp. 2184–2185, Dec. 1999.

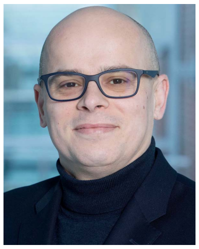
[51] C. She *et al.*, "A tutorial on ultra-reliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," Sep. 2020, *arXiv:2009.06010*. [Online]. Available: http://arxiv.org/abs/2009.06010

[52] M. Angjelichinoski, K. F. Trillingsgaard, and P. Popovski, "A statistical learning approach to ultra-reliable low latency communication," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5153–5166, Jul. 2019.

Federico Chiariotti (Member, IEEE) received the bachelor's and master's degrees *(cum laude)* in telecommunication engineering from the University of Padova, in 2013 and 2015, respectively, and the Ph.D. degree in information engineering from the University of Padova, Italy, in 2019. He is currently a Post-Doctoral Researcher with the Department of Electronic Systems, Aalborg University, Denmark. He has authored over 30 published papers on wireless networks and the use of artificial intelligence techniques to improve their performance. He was a recipient of the Best Paper Award at several conferences, including the IEEE INFOCOM 2020 WCNEE Workshop. His current research interests include network applications of machine learning, transport layer protocols, smart cities, bike sharing system optimization, and adaptive video streaming.

Olga Vikhrova received the master's and master's degrees in computer science and information technologies from RUDN University, Russia, in 2014 and 2017, respectively. She is currently pursuing the Ph.D. degree with the Information Engineering Department, University Mediterranea of Reggio Calabria, Italy. In 2020, she was a Visiting Ph.D. Student with the Department of Electronic Systems, Aalborg University, Denmark. Her current research interests include cellular IoT networks, distributed edge computing, multicast communications, and non-terrestrial networks.

Beatriz Soret (Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications from the Universidad de Malaga, Spain, in 2002 and 2010, respectively. She is currently an Associate Professor with the Department of Electronic Systems, Aalborg University, Denmark. Before, she has been with Nokia Bell-Labs and GomSpace. She has coauthored more than 60 publications in journals and conference proceedings, 16 patents in the area of wireless communications, and she received the Best Paper Award in IEEE Globecom 2013. Her research interests are within IoT connectivity, 5G and post-5G systems, non-terrestrial networks, low-latency, and high reliable communications and timing in communications.



Petar Popovski (Fellow, IEEE) received the Dipl.-Ing. and M.Sc. degrees in communication engineering from the University of Sts. Cyril and Methodius in Skopje and the Ph.D. degree from Aalborg University, in 2005. He is currently a Professor with Aalborg University, where he heads the section on Connectivity. He received an ERC Consolidator Grant (2015), the Danish Elite Researcher award (2016), IEEE Fred W. Ellersick prize (2016), IEEE Stephen O. Rice prize (2018), Technical Achievement Award from the IEEE Technical Committee on Smart Grid Communications (2019), and the Danish Telecommunication Prize (2020). He is a Member at Large at the Board of Governors in IEEE Communication Society, a Steering Committee Member of IEEE Communication Theory Workshop, IEEE SmartGridComm, and IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING. He is currently an Area Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was the General Chair for IEEE SmartGridComm 2018 and IEEE Communication Theory Workshop 2019. His research interests are in the areas of wireless communication and communication theory. He authored the book *Wireless Connectivity: An Intuitive and Fundamental Guide*, published by Wiley in 2020.