

Latency and Peak Age of Information in Non-Preemptive Multipath Communications

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

Abstract—Multipath communication is a critical technology to provide Quality of Service (QoS) to interactive and Internet of Things (IoT) monitoring and control applications. In this work, we model the exemplary case with two paths and consider different strategies that exploit redundancy and coding to improve the timing performance of wireless communications. We consider two disparate scenarios, in which the data blocks are generated via a Markovian and a deterministic process, respectively. We consider simple scheduling and coding schemes, considering both lossless and lossy encoding, and modeling the resulting process as a fork-join queue with different arrival processes. We analyze the full distribution of two relevant metrics for the two-path case: the packet delay and the Peak Age of Information (PAoI), which measures the freshness of the information at the receiver. The results show interesting trade-offs between the update frequency, latency, PAoI, and level of compression, with interesting implications for system designers.

Index Terms—Multipath communications, queuing theory, Age of Information, fork-join queues

I. INTRODUCTION

Over the past decade, the traditional view of latency as the only significant timing metric in communications has gradually eroded: Age of Information (AoI) has attracted a significant interest in the research community since its inception in 2012 [1], as it can better represent the delay perceived by users in real-time applications. AoI does not just measure the time between when a block of data is generated and when it is delivered, but keeps increasing until the next block has been delivered. As the name suggests, the AoI is the *age* of the information that the receiver has at any moment, and intuitively, it is more useful for control tasks, as it represents the actual delay between the state of the system and the controller's action: the latter will be based on old information, and the older the representation of the state of the system, the worse the control performance will be. In particular, the Peak Age of Information (PAoI), which measures the age right before the next block is delivered, is a good proxy for the worst-case timing discrepancy of a human or automatic controller.

The first and foremost use case for AoI is Internet of Things (IoT) monitoring, which has been studied extensively, but Virtual Reality (VR) and other interactive applications might also benefit from this metric. For example, AoI for VR

TABLE I: Main derived results for each considered scheme.

Scheme	Latency ($D/M/2$)	Latency ($M/M/2$)	PAoI ($D/M/2$)	PAoI ($M/M/2$)
Alternating	Exact	Exact	Lower bound	Simulation
Split	Exact	Exact	Exact	Exact
Replicated	Exact	Exact	Lower bound	Exact
Coded	Exact	Exact	Lower bound	Exact

content is crucial to enable the *digital twin* concept, which enables remote inspection and operation of cyber-physical systems, and for human-in-the-loop control, in which parts of a manufacturing system are automated, while other parts are controlled directly by a human operator [2], [3]. However, these applications can suffer from significant timeliness-related issues on wireless networks [4]: as the sense of presence and immersion is critical, and the size of each omnidirectional frame can be huge, this puts a strain on the notoriously volatile wireless links, with a high risk of congestion and sudden delay increases, which are perceived by the user as an annoying loss of smoothness in the VR experience. Adaptive video content can reduce delay by compressing the video, trading video quality for a smaller, more predictable delay [5], but the unpredictability of the wireless propagation environment can make this task very complex.

In this context, multipath communications [6] can be a way to provide Quality of Service (QoS) to both IoT monitoring and VR services. This work models such a communication scenario as a theoretical fork-join system, which can handle either periodic updates (as is the case for VR, which often operates at a constant frame rate) or Poisson updates (as is most common for sensor monitoring scenarios in the IoT). We focus on the case with only two paths, and analytically derive the complete distribution of the latency and PAoI, or bounds for these distributions, with four different transmission schemes. In particular, we consider schemes that try to reduce the load on the multipath connection by splitting the traffic between the two paths, as well as schemes that exploit redundancy across both paths to protect the transmission from the unpredictable errors and delays on any individual path, and derive the exact distribution of the latency and PAoI, and a lower bound to them for others, as detailed in Table I. We draw some considerations on the performance of each scheme, finding interesting results in the trade-off between reliability (i.e., delivering as many data blocks as possible in error-prone scenarios), update frequency and quality, and latency or age. Part of this analysis, including only one scenario and one strategy among the ones in this paper, was presented as a conference publication [7].

The rest of the paper is organized as follows: Sec. II presents

Federico Chiariotti (corresponding author, fchi@es.aau.dk) and Petar Popovski (petarp@es.aau.dk) are with the Department of Electronic Systems, Aalborg University, Denmark. Beatriz Soret (bsoret@ic.uma.es) is with the Telecommunication Research Institute (TELMA), Universidad de Málaga, Spain. This work was partly funded by the IntelliIoT project under the H2020 framework grant ID 957218, and partly by the Villum Investigator Grant “WATER” from the Velux Foundation, Denmark.

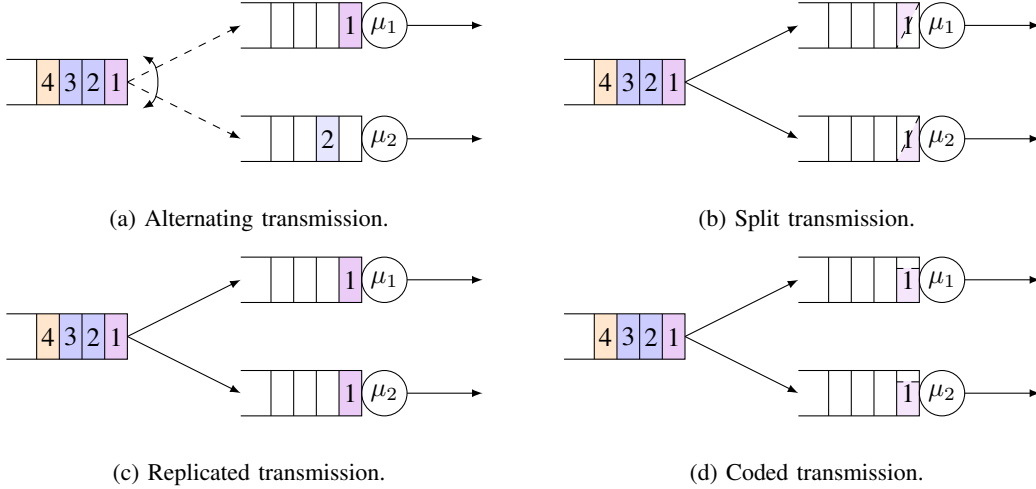


Fig. 1: Depiction of the four transmission schemes.

an overview of the related work on multipath communications, queuing models, and AoI. Our overall system model is described in Sec. III, and the analyses for the systems with deterministic and Poisson updates are described in Sec. IV and Sec. V, respectively. Sec. VI then presents our simulation settings and results, and Sec. VII concludes the paper and presents some avenues of future work.

II. RELATED WORK

Since the concept of AoI was introduced in the early 2010s [1], it has been the subject of intense study, being the preferred latency metric for real-time control and context-aware IoT applications [8]. In these applications, the end receiver is interested in a fresh knowledge of the remotely controlled system, rather than the packet delay. Most theoretical results derive the average AoI or the PAoI. Much smaller is the number of works deriving higher moments or the full distribution of the age (see e.g. [6] [9]), although a reliable system design requires knowing the probability of occurrence of rare, but extremely damaging events.

Initial studies focused on simple queuing systems with a single node and unicast scheduled transmissions [10], [11]. However, there is already a significant number of papers addressing more complex communication scenarios and topologies, such as models for random access and ALOHA [12] [13], multicast and broadcast [14], or multi-hop transmission [15], [16]. More complex network models, in which multiple nodes transmit updates to each other and act as relays, are another interesting scenario, in which it is possible to analyze the spatial diffusion of information [17] or the age at the last node [18].

Our system model is based on a fork-join queue [19], where incoming tasks are split into several servers and joined again before departing the system. This model has been used for parallel multitasking in computation and communications networks [20]. Most of the works assume that, at any time instant, tasks can be canceled and abandon their respective queue. [21] studies path redundancy in the context of cloud systems with a fork-join model, to understand the trade-off latency-computing cost. [22] analyzes the transmission of

redundant requests to multiple servers for a faster execution in terms of average latency, at the cost of increased system load. It is observed that not having redundancy is optimal for highly loaded systems if service times are memoryless. The authors in [23] present a study on delay-optimal scheduling of replications in centralized and distributed multi-server systems.

There has also been some work on bounds to worst-case performance in fork-join queues, mostly concerned with the tail of the latency distribution [24]. These delay bounds concern themselves with Markovian generation processes [25], as they model parallel or distributed processing of data generated at random times. To the best of our knowledge, the question of delay bounds with periodic traffic, such as video or VR frames, has not been investigated in either the communication or distributed processing literature.

Other works in the literature consider the age-distortion trade-off in the case of imperfect coding [26], defining the compression rate at the sender side to trade accuracy of the information for its freshness [27]. It is also possible to explicitly include considerations on the age in the encoding procedure [28], taking a further step towards the joint consideration of age and compression. As our focus is on the multipath transmission strategy, we consider a simpler coded policy in which the compression is constant and decided beforehand.

Almost none of the fork-join queuing works in the literature analyzes the AoI. To the best of our knowledge, the only exceptions are [29], which gives the average age for distributed computing with K over N coding, and [30], which addresses the trade-off between average age and average delay considering the scheduler and routing (selected path) of an M -server system with Poisson generation. They prove that a system designed to minimize the age will do it at the expenses of high waiting times and service times for the packets that do not contribute to the age metric (the *non-informative* packets) and the average and variance of the packet delay will therefore increase enormously.

Unlike most of the fork-related works, we do not consider

removal of task in the buffers, since age-sensitive applications will typically have no feedback and there is no way for the transmitter to know when the replicas/pieces of the data block have been delivered to any of the receivers. Even if feedback is available, feedback may be delayed, and this factor should not be ignored, as it can be a significant factor in communication systems. Assuming no feedback provides a consistent measure of timing, which only depends on the reception delay of the actual information. In the same way, we do not consider packet preemption, but use a First Come First Serve (FCFS) policy for all queues, as we consider the packet delivery ratio to be an important parameter even if it increases the age.

III. SYSTEM MODEL

We consider a wireless multipath system, in which information blocks of constant size are generated following a random process. The generated blocks are then delivered to a user through a multipath connection using two different Radio Access Technologies (RATs). We model the two paths as two separate queuing systems with Markovian service. The two paths, dubbed 1 and 2, then have exponential service times, with average rates μ_1 and μ_2 , respectively: the service time is exponentially distributed to account for physical and medium access issues, which introduce a significant volatility in the delivery of the data. We also consider them to have infinite queues with FCFS service. The AoI literature often uses preemptive schemes instead of infinite-buffer FCFS, as system designers are mostly concerned with getting the most recent update and use preemptive systems that can discard older packets once newer ones arrive in the queue, but we consider reliability and delivery of even older information as an important system parameter, even if it comes at a slight cost in terms of AoI. In particular, VR applications would require the use of FCFS for efficient compression, as video compression is often based on the difference between subsequent frames, and discarding frames or delivering them out of order could be harmful for the decoder.

In the following, we will refer to the individual paths in the connection as “paths” or “channels” interchangeably, using “system” to indicate the overall multipath connection. We will consider both deterministic and Markovian generation processes, indicating the block generation frequency as λ in both cases. The overall system is then a $D/M/2$ or $M/M/2$ fork-join queue, as the two paths have separate queues and the transmitter has a scheduler, i.e., it can choose which connection to send the data through. We assume that the two paths also function as Packet Erasure Channels (PECs), with erasure rates ε_1 and ε_2 , respectively. We define the load ρ_j on each path j as:

$$\rho_j = \frac{L\lambda_j}{\mu_j}, \quad j = 1, 2, \quad (1)$$

where λ_j is the average generation rate on the path, which depends on the scheduling process, and L is the normalized size of the packets: an uncoded block is considered to have $L = 1$. We consider four different scheduling strategies, which are depicted in Fig. 1 and described for the general case with M paths:

TABLE II: Main notations used in the paper.

Symbol	Meaning	Symbol	Meaning
μ_j	Service rate of path j	ε_j	Error probability for path j
λ_j	Average generation rate on path j	L	Length of a packet
λ	Data block generation rate	η	Coding rate
$Q_{i,j}$	Queue state for block i at path j	g_i	Generation time of block i
r_i	Delivery time for block i	T_i	Latency for block i
$\Theta(t)$	AoI at time t	Δ_i	PAoI at block i
τ	Average inter-generation time	σ_j	$D/M/1$ steady-state parameter for path j
ρ_j	Load on path j	G_i	Relevance of the i -th packet
F_i	Number of consecutive failures before i	$x_{i,j}$	Transmission success indicator
$\xi_j(q, t)$	Auxiliary integral function	ϕ_j	Probability of receiving only on link j
$\phi_{1,2}$	Probability of receiving both packets		

- The *alternating* transmission scheme, shown in Fig. 1a, is a simple round-robin scheduler: each block corresponds to a packet, which is sent over the selected path. In this case, the multipath connection is used to reduce the load on the paths: the generation rate is $\lambda_j = \frac{\lambda}{M}$ for any connection, while we have $L = 1$.
- The *split* transmission scheme, shown in Fig. 1b, divides the block into M packets of size $\frac{1}{M}$, each of which is sent over a different path. In this case, we have $\lambda_j = \lambda$, but $L = \frac{1}{M}$, so the overall load on the system is the same.
- The *replicated* transmission scheme, shown in Fig. 1c, replicates each block M times, generating M packets which are sent over all paths at once. Naturally, this increases the load on the system, as we have $\lambda_j = \lambda$ on all paths and $L = 1$, but it also provides error protection, as up to $M - 1$ erasures can be compensated by the remaining replicas.
- The *coded* transmission scheme, shown in Fig. 1d, is a hybrid between the replicated and split schemes: the transmitter sends two packets of equal size $\eta \in [\frac{1}{M}, 1]$ on each path. We consider a lossy compression scheme with orthogonal information, such as, e.g., Multiple Description Coding (MDC) for video content, so that each packet is decodable individually, and provides a lower-quality representation of the information in the block. If both packets are delivered, the block can be decoded at the full quality. The quality and size of each individual packet depend on the coding rate η : a value $\eta = \frac{1}{M}$ corresponds to the replicated scheme, while $\eta = 1$ corresponds to the split scheme. The generation rate at each path is still $\lambda_j = \lambda$, while we have $L = \frac{1}{M\eta}$, as the block is encoded and then split over the M paths.

In all schemes but the first one, packet generation on the paths is synchronized, i.e., one packet is sent on each path for each block. Naturally, more intelligent schedulers are possible: a more advanced version of the alternating scheme would schedule the block based on the occupancy of each queue, always sending packets on the less congested path. However, this kind of scheduler is difficult to study analytically, and we will only provide simulation results for it. We can also think of an adaptive coded scheme, which changes the compression rate depending on the state of the two queues, but this is beyond the scope of this work, where we aim to study the basic features of the four alternatives from Fig. 1 in the simpler case with $M = 2$. Furthermore, our model assumes that each path

has independent service times for analytical tractability: the assumption is verified when using different communication technologies with uncorrelated service and cross-traffic. If paths are correlated, the derivation of latency and PAoI are much more complex, and beyond the scope of our work.

In the following, we will represent Random Variables (RVs) using capital letters, and their values with lower-case letters. Vectors are represented in bold, and p_X represents the Probability Density Function (PDF) of RV X , while P_X represents its Cumulative Density Function (CDF). We define the state of the overall system when the i -th block is generated as $\mathbf{Q}_i = (Q_{i,1}, Q_{i,2})$, where $Q_{i,j}$ represents the number of packets in channel j , including the one in service. We derive next the system time and peak age distributions for the four transmission schemes, first for the alternating scheme, then for the coding-based ones. The main notations used in the following are listed in Table II. We also remark that the *lower bound* to a random variable in the usual stochastic sense corresponds to an *upper bound* on its CDF: as the computed probability that the variable will be lower than a certain value is always higher than its actual CDF for that value, any quantile of the actual distribution will be higher than the corresponding quantile for the bound.

We also give a definition of the latency, AoI, and PAoI. We consider the i -th generated block, which is generated at time g_i . The block is then transmitted through the fork-join system, and decoded at time r_i ; if the block is not received correctly due to erasures, we consider $r_i = \infty$. The latency T_i is then simply given by $r_i - g_i$. The definition of the AoI $\Theta(t)$ is also simple, as it represents the time elapsed since the most recent correctly received block was generated:

$$\Theta(t) = t - \max_{i \in \mathbb{N}: r_i \leq t} g_i. \quad (2)$$

The PAoI is a sampling of the AoI right before a new informative update is received:

$$\Delta_i = r_i - \max_{j < i: r_j \leq r_i} g_j. \quad (3)$$

It is only relevant for blocks that contain new information, i.e., if $\nexists j > i : r_j \leq r_i$.

We consider the high-quality and low-quality versions of the blocks delivered by the coded scheme separately, as any system mixing the two would necessarily require a definition of *when* to use each. As such, our results will distinguish between high-quality, or HQ, age and latency, and low-quality, or LQ. We can think of a deadline-based system, which would decode the data using only one packet if the other does not arrive within a given time or if the AoI passes a threshold value, but defining such a scheme is outside the scope of this work, as it would require a more extensive knowledge of the application. The coded scheme is also useful for multicast scenarios, as multiple receivers may have different priorities and use different schemes to decide when to decode each data block.

IV. ANALYSIS OF THE $D/M/2$ SYSTEM

We first consider a $D/M/2$ system, in which blocks of data are generated at a constant and known interval $\tau = \frac{1}{\lambda}$. We

will now derive the PAoI and latency distribution for the four transmission strategies, providing bounds in the cases in which analytical derivation is not possible. The coded transmission strategy is a combination of the split and replicated ones with a different η , but it has two different quality levels for the received information. The PDFs of the system time and PAoI are the same as for the split transmission for the high-quality version, and the same as the replicated transmission for the lower-quality version.

A. Alternating transmission

In this scheme, packets are divided between the two paths in a round-robin fashion. Each of the two paths is independent from the other and can be treated individually, with a packet generation process with only half the rate. We can then consider each path as a $D/M/1$ system, in which the inter-generation period for each path is 2τ . We then have the following distribution of the state $q_{i,j}$ just before a new packet i is generated, following the well-known formula derived by Erlang [31]:

$$p_{Q_{i,j}}(q_{i,j}) = (1 - \sigma_j) \sigma_j^{q_{i,j}}, \quad (4)$$

where the parameter σ_j is the steady-state distribution parameter, which corresponds to the only solution in $(0, 1)$ to

$$x = e^{\mu_j \tau (x-1)}. \quad (5)$$

The existence of σ_j is guaranteed if path j is stable, i.e., if $\rho_j = \frac{1}{2\mu_j \tau} < 1$ $j \in \{1, 2\}$. If we know the size of the queue that packet i finds on path j , the system time $T_{i,j}$ is Erlang distributed, with parameters $q_{i,j}$ and μ_j . Since the two paths are independent, we can compute the PDF of $T_{i,j}$ by applying the law of total probability, using the steady-state distribution from (4):

$$\begin{aligned} p_{T_{i,j}}(t) &= (1 - \varepsilon_j) \sum_{q_{i,j}=0}^{\infty} (1 - \sigma_j) \sigma_j^{q_{i,j}} \frac{\mu_j^{q_{i,j}+1} t^{q_{i,j}} e^{-\mu_j t}}{q_{i,j}!} u(t) \\ &= (1 - \varepsilon_j) \mu_j (1 - \sigma_j) e^{-\mu_j (1 - \sigma_j) t} u(t), \end{aligned} \quad (6)$$

where $u(t)$ is the step function, equal to 1 if $t \geq 0$ and 0 otherwise. It is also easy to derive the CDF $P_{T_{i,j}}(t)$. The overall latency distribution, without a condition on the packet index, is then simply:

$$p_{T_i}(t) = \frac{p_{T_{i,1}}(t) + p_{T_{i,2}}(t)}{2}. \quad (7)$$

If we consider error-prone channels, the exact PAoI distribution is hard to compute, as packets might be delivered out of order: if an older packet is delivered after a newer one, it is considered as irrelevant. We then define the relevance of packet i as a Bernoulli RV G_i , which is 1 if the packet is relevant and 0 otherwise:

$$G_i = \begin{cases} 0, & \text{if } \exists \ell > i : r_\ell > r_i; \\ 1, & \text{otherwise.} \end{cases} \quad (8)$$

Theoretically, there might be any number of consecutive reordered packets in an error-prone system, while we never have more than 1 packet delivered out of order in error-free

systems (as the third packet is blocked by the first one, which has not been delivered yet). In the following, we then compute a lower bound to the PAoI, assuming that, for the transmission of packet i , packet $i-3$ has already been delivered on the other path, i.e., multiple packets can be lost, but we never have more than 2 consecutive packets are delivered out of order. As the condition is always true in the error-free case, this bound is exact in those conditions. If i is odd, the probability $p_{G_{i,1}}(1)$ that packet i is relevant is then given by the probability that packet $i+1$ is delivered after it:

$$\begin{aligned} p_{G_{i,1}}(1) &= P_{T_{i,1}}(\tau) + \int_{\tau}^{\infty} p_{T_{i,1}}(t)(1 - P_{T_{i+1,2}}(t - \tau))dt \\ &= (1 - \varepsilon_1) \left(1 - \frac{(1 - \varepsilon_2)\mu_2 e^{-\mu_1(1-\sigma_1)\tau}}{\sum_{j=1}^2 (1 - \sigma_j)\mu_j} \right). \end{aligned} \quad (9)$$

The conditioned PDF $p_{T_{i,1}|G_{i,1}}(t|1)$ can then be computed by applying Bayes' theorem:

$$p_{T_{i,1}|G_{i,1}}(t|1) = \frac{[1 - u(t - \tau)(1 - P_{T_{i+1,2}}(t - \tau))] p_{T_{i,1}}(t)}{p_{G_{i,1}}(1)}. \quad (10)$$

We denote the number of consecutive failures before packet i as F_i . We then have the following bound if there is at least one failure:

$$P_{\Delta_{i,1}|F_i}(\delta|f) \leq P_{T_{i,1}|G_{i,1}}(\delta - (f+1)\tau|1) \forall f \geq 1. \quad (11)$$

If there are no failures, i.e., $F_i = 0$, we need to consider whether the most recent received packet before i is number $i-1$ (over link 2) or $i-2$ (over link 1):

$$\begin{aligned} P_{\Delta_{i,1}|F_i}(\delta|0) &\leq \frac{P_{T_{i-1,2}}(\delta)P_{T_{i,1}|G_{i,1}}(\delta - \tau|1)}{1 - \varepsilon_2} \\ &\quad + \left(1 - \frac{P_{T_{i-1,2}}(\delta)}{1 - \varepsilon_2} \right) P_{T_{i,1}|G_{i,1}}(\delta - 2\tau|1). \end{aligned} \quad (12)$$

We can then use the law of total probability to remove the condition on F :

$$P_{\Delta_{i,1}}(\delta) \leq \sum_{f=0}^{\lfloor \frac{\delta}{\tau} \rfloor} \varepsilon_1^{\lfloor f/2 \rfloor} \varepsilon_2^{\lfloor f/2 \rfloor} (1 - \varepsilon_{2-\text{mod}(f,2)}) P_{\Delta_{i,1}|F_i}(\delta|f), \quad (13)$$

where $\text{mod}(m, n)$ is the integer modulo function. The same is true for the second path, after swapping the indices. We can then get the overall age bound CDF:

$$P_{\Delta_i}(\delta) \leq (1 - \varepsilon_1)p_{G_{i,1}}(1)p_{\Delta_{i,1}}(\delta) + (1 - \varepsilon_2)p_{G_{i,2}}(1)p_{\Delta_{i,2}}(\delta). \quad (14)$$

As we remarked above, the bound is exact if $\varepsilon_1 = \varepsilon_2 = 0$.

We can also easily give a bound to the average AoI in the case in which $\varepsilon_1 = \varepsilon_2 = \varepsilon$. We know from [1] that the average AoI can be computed geometrically as:

$$\mathbb{E}[\Theta] = \mathbb{E}[TY] + \frac{\mathbb{E}[Y^2]}{2}. \quad (15)$$

In this case, the inter-generation time Y between subsequent valid updates is just τ multiplied by a geometric distribution with parameter $1 - \varepsilon$, and the system time is independent

from the valid update inter-generation time, resulting in the following bound:

$$\mathbb{E}[\Theta] \geq \left(\frac{\tau}{2(1 - \varepsilon)} + \sum_{j=1}^2 \frac{p_{G_j}(1)\mu_j^{-1}(1 - \sigma_j)^{-1}}{(p_{G_1}(1) + p_{G_2}(1))} \right) \frac{\tau}{1 - \varepsilon}. \quad (16)$$

B. Split transmission

We now consider a system in which smaller packets are sent through the system, with coding rate $\eta = 1$. The steady-state distribution for this kind of system was derived in [32] following Palm probability theory [33]. The joint distribution of \mathbf{Q}_i , considering the overall system just before packet i is generated, is given by:

$$p_{\mathbf{Q}_i}(\mathbf{q}_i) = (1 - \sigma_1)(1 - \sigma_2)\sigma_1^{q_{i,1}}\sigma_2^{q_{i,2}}, \quad (17)$$

where σ_j is defined in a similar way to (5), including the coding:

$$x = e^{2\eta\mu_j\tau(x-1)}. \quad (18)$$

In the following, we will use the notation $\mu'_j = 2\eta\mu_j$ for the sake of brevity. We now consider the i -th generated block, which causes two packets to be generated simultaneously at both paths at time g_i . The two packets will then be delivered to the receiver at times $r_{i,1}$ and $r_{i,2}$, depending on the system time of the two paths. We denote the success of the transmission on path j as the Bernoulli RV $\chi_{i,j}$, equal to 1 if the packet is transmitted successfully and 0 otherwise.

We can then define the system time T_i^{\max} , which is equivalent to the delay between the generation of the block and the reception of both packets:

$$T_i^{\max} = \max_{j \in \{1,2\}} (r_{i,j} - g_i)\chi_{i,j}. \quad (19)$$

The distribution of T_i^{\max} is the maximum of the two distributions of the system times at the two paths, which are independent if the state of the two queues is given. In the following, we omit the index of the block i for brevity. We also define the auxiliary term $\xi_j(q, t)$ as:

$$\xi_j(q, t) = \sum_{n=0}^q \frac{(\mu'_j t)^n e^{-\mu'_j t}}{n!}. \quad (20)$$

If $q = 0$, we simply have $\xi_j(0, t) = e^{-\mu'_j t}$. If we know the values of Q_1 and Q_2 , we then get the following distribution of the system time, corresponding to the maximum between two Erlang-distributed variables, which correspond to the system time for the two paths:

$$\begin{aligned} p_{T^{\max}|\mathbf{Q}}(t|(q_1, q_2)) &= p_s^{\max} u(t) \left[\frac{(1 - \xi_1(q_1, t))(\mu'_2 t)^{q_2+1}}{te^{\mu'_2 t} q_2!} \right. \\ &\quad \left. + \frac{(1 - \xi_2(q_2, t))(\mu'_1 t)^{q_1+1}}{te^{\mu'_1 t} q_1!} \right]. \end{aligned} \quad (21)$$

We can then remove the condition on Q_2 from the system time distribution, using the law of total probability to condition only on the first queue's state:

$$p_{T^{\max}|Q_1}(t|q_1) = (1 - \sigma_2) \left[(1 - \xi_1(q_1, t)) \mu'_2 e^{-\mu'_2(1-\sigma_2)t} + \frac{\mu'_1(\mu'_1 t)^{q_1} (1 - e^{-\mu'_2(1-\sigma_2)t})}{(1 - \sigma_2) e^{\mu'_1 t} q_1!} \right] p_s^{\max} u(t). \quad (22)$$

We can finally remove the condition on the state of the first queue in the same way:

$$p_{T^{\max}}(t) = \sum_{j=1}^2 (1 - \sigma_j) \mu'_j e^{-\mu'_j(1-\sigma_j)t} (1 - e^{-\mu'_{3-j}(1-\sigma_{3-j})t}) \times p_s^{\max} u(t). \quad (23)$$

The maximum system time represents the latency for a system which requires both packets to fulfill the request. Network coded transmissions, in which both packets contain part of the information required at the receiver, is one example of this kind of system. The CDF of the system time is:

$$P_{T^{\max}}(t) = [1 - e^{-\mu'_1(1-\sigma_1)t} - e^{-\mu'_2(1-\sigma_2)t} + e^{-2\eta(\mu'_1(1-\sigma_1) + \mu'_2(1-\sigma_2))t}] p_s^{\max} u(t). \quad (24)$$

We can now consider the distribution of the PAoI, denoted by Δ . In this case, we need to consider the possibility of block failures. We denote the number of failures as F : if there are f consecutive failures, the PDF of the PAoI is simply given by:

$$P_{\Delta|F}(\delta|f) = P_{T^{\max}}(\delta - (f+1)\tau) u(\delta - (f+1)\tau) \quad \forall f \geq 1. \quad (25)$$

This is due to the deterministic nature of the packet generation process, which increases the age by τ for every failure. As failures are independently distributed and the success probability is p_s^{\max} , we can now apply the law of total probability to remove the condition:

$$P_{\Delta}(\delta) = p_s^{\max} \sum_{f=0}^{\lfloor \frac{\delta}{\tau} \rfloor - 1} (1 - p_s^{\max})^f P_{T^{\max}}(\delta - (f+1)\tau). \quad (26)$$

We can get the average AoI in the same way as for the alternating scheme:

$$\mathbb{E}[\Theta] = \frac{\tau}{p_s^{\max}} \left(\frac{\mu'_1(1-\sigma_1) + \mu'_2(1-\sigma_2)}{\mu'_1\mu'_2(1-\sigma_1)(1-\sigma_2)} - 1 + \frac{\tau}{2p_s^{\max}} \right). \quad (27)$$

C. Replicated transmission

We can now examine the replicated transmission scheme, which has $\lambda_j = \lambda$ and $\eta = 0.5$. The state of the two queues is still independent and given by (17), substituting the correct value of η . We then have a probability $p_s^{\min} = 1 - \varepsilon_1\varepsilon_2$ of fulfilling each request. Furthermore, we define $\phi_{1,2} = (1 - \varepsilon_1)(1 - \varepsilon_2)$, $\phi_1 = (1 - \varepsilon_1)\varepsilon_2$, and $\phi_2 = (1 - \varepsilon_2)\varepsilon_1$, to identify the probability of having a block with two received packets, only the packet on the first path being received, and only the packet on the second path being received, respectively.

Since both packets contain the same information, the latency for successful blocks in this scheme is the minimum system time T_i^{\min} , i.e., the system time of the first packet to be received:

$$T_i^{\min} = \min_{j \in \{1,2\}; \chi_{i,j}=1} r_{i,j} - g_i. \quad (28)$$

The minimum system time is then a RV whose distribution is the minimum between the distributions of the two independent paths. In the following, we omit the index of the packet i for the sake of readability. We can now compute the conditioned PDF of this distribution, considering the system state \mathbf{Q} as known. The system time is now the minimum between two Erlang distributed RVs, whose distribution is given by:

$$p_{T^{\min}|\mathbf{Q}}(t|(q_1, q_2)) = u(t) \sum_{j=1}^2 \frac{(\phi_j + \phi_{1,2}\xi_{3-j}(q_{3-j}, t))}{te^{\mu'_j t} (\mu'_j t)^{-(q_j+1)} q_j!}. \quad (29)$$

As we did for the split case, we now apply the law of total probability to remove the conditions on Q_1 and Q_2 :

$$\begin{aligned} p_{T^{\min}}(t) &= \sum_{q_1=0}^{\infty} \sum_{q_2=0}^{\infty} p_{T^{\min}|\mathbf{Q}}(t|(q_1, q_2)) \prod_{j=1}^2 (1 - \sigma_j) \sigma_j^{q_j} \\ &= u(t) \sum_{j=1}^2 \left[\phi_j (1 - \sigma_j) \mu'_j e^{-(1-\sigma_j)\mu'_j t} \right. \\ &\quad \left. + \phi_{1,2} \sum_{q_{3-j}=0}^{\infty} (1 - \sigma_{3-j}) \sigma_{3-j}^{q_{3-j}} \sum_{n=0}^{q_{3-j}} \frac{(\mu'_{3-j} t)^n}{n! e^{\mu'_{3-j} t}} \right]. \end{aligned} \quad (30)$$

As the factor σ is guaranteed to be in the open interval $(0, 1)$ due to the stability condition on the two paths, its geometric series converges and the inversion of the two summations in the second term is possible. We solve the series for link j :

$$\begin{aligned} \sum_{q_j=0}^{\infty} (1 - \sigma_j) \sigma_j^{q_j} \sum_{n=0}^{q_j} \frac{(\mu'_j t)^n}{n!} &= (1 - \sigma_j) \sum_{n=0}^{\infty} \frac{(\mu'_j t)^n}{n!} \sum_{q=n}^{\infty} \sigma_j^{q_j} \\ &= \sum_{n=0}^{\infty} \frac{(\sigma_j \mu'_j t)^n}{n!} = e^{\sigma_j \mu'_j t}. \end{aligned} \quad (31)$$

We can then give the unconditioned PDF:

$$\begin{aligned} p_{T^{\min}}(t) &= u(t) \left[\sum_{j=1}^2 \phi_j (1 - \sigma_j) \mu'_j e^{-\mu'_j(1-\sigma_j)t} \right. \\ &\quad \left. + \phi_{1,2} (\mu'_1 + \mu'_2) e^{-(\mu'_1(1-\sigma_1) + \mu'_2(1-\sigma_2))t} \right]. \end{aligned} \quad (32)$$

We can also get the CDF of the latency:

$$\begin{aligned} p_{T^{\min}}(t) &= u(t) \left[\sum_{j=1}^2 \left(\phi_j (1 - e^{-\mu'_j(1-\sigma_j)t}) \right) \right. \\ &\quad \left. + \phi_{1,2} \left(1 - e^{-(\sum_{j=1}^2 \mu'_j(1-\sigma_j)t)} \right) \prod_{j=1}^2 (1 - \sigma_j) \right]. \end{aligned} \quad (33)$$

The CDF of the PAoI can be derived in the same way as for the split scheme. However, in this case, the possibility of

reordered packets makes the derived value a lower bound of the actual PAoI. This bound does not consider the case in which a packet from block $i + 1$ is delivered before the first one from packet i , which is possible if one path is lossy and the other is highly congested:

$$P_{\Delta}(\delta) \leq \sum_{f=0}^{\lfloor \frac{\delta}{\tau} \rfloor - 1} (1 - p_s^{\min})^f P_{T^{\min}}(\delta - (f + 1)\tau). \quad (34)$$

The bound is exact if $\varepsilon_1 = \varepsilon_2 = 0$, as in that case block $i + 1$ is never delivered before block i .

We can also give a lower bound to the average AoI, as we did for the other schemes:

$$\mathbb{E}[\Theta] \geq \frac{\tau}{(p_s^{\min})^2} \left(\frac{\phi_{1,2}}{\prod_{j=1}^2 \mu'_j(1 - \sigma_j)} + \sum_{j=1}^2 \frac{\phi_j}{\mu'_j(1 - \sigma_j)} + \frac{\tau}{2} \right). \quad (35)$$

D. Coded transmission

As the coded transmission is simply a combination of the previous two cases, we can give the latency for the low- and high-quality versions directly. The latency for the high-quality version is given by (21), while the latency for the low-quality version is given by (32), using the appropriate value of η . The same goes for the PAoI, which is given by (26) for the high-quality version and lower-bounded by (26) for the low-quality version.

V. ANALYSIS OF THE $M/M/2$ SYSTEM

We now repeat the analysis for the $M/M/2$ fork-join system, in which blocks of data arrive according to a Poisson process with rate λ , which has an average inter-arrival time $\tau = \frac{1}{\lambda}$. We will now derive the PAoI and latency distribution for the four transmission strategies, providing bounds in the cases in which analytical derivation is not possible.

The derivation of the bounds on the latency and PAoI is supported by the definition of several auxiliary definitions that simplify the notation. First, the parameter $\psi_{i,j}$ is given by:

$$\psi_{i,j} = q_{i-1,j} - q_{i,j} + 1. \quad (36)$$

We also define $\psi_i = \psi_{i,1} + \psi_{i,2}$. Then we define the auxiliary function $\theta_{m,n}^{(\lambda)}(a)$:

$$\begin{aligned} \theta_{m,n}^{(\lambda)}(a) &= \int_0^a e^{\lambda x} x^m (a - x)^n dx \\ &= \sum_{i=0}^n \binom{n}{i} \frac{(m+i)! a^{n-i}}{\lambda^{m+i+1} (-1)^m} \left(e^{\lambda a} \sum_{j=0}^{m+i} \frac{(-\lambda a)^j}{j!} - 1 \right), \quad \lambda \neq 0. \end{aligned} \quad (37)$$

If $\lambda = 0$, the value is simpler to derive:

$$\theta_{m,n}^{(0)}(a) = \int_0^a x^m (a - x)^n dx = \binom{m+n}{m} \frac{a^{m+n+1}}{m+n+1}. \quad (38)$$

Another solution for negative values of λ is given in [34], using the Beta function $\beta(x, y)$ and the confluent hypergeometric function ${}_1F_1(a, b, z)$:

$$\begin{aligned} \theta_{m,n}^{(\lambda)}(a) &= \beta(m+1, n+1) {}_1F_1(m+1, m+n+1, -\lambda a) \\ &\quad \times a^{m+n+1}, \quad \lambda < 0. \end{aligned} \quad (39)$$

As this solution avoids summing large values with alternating signs, it is numerically more stable.

A. Alternating transmission

In this case, the alternated transmission scheme corresponds to two independent FCFS $G/M/1$ queues, in which the arrival distribution is $\text{Erl}(2, \lambda)$. This case has been analyzed in [35] for a general $\text{Erl}(k, \lambda)$ distribution, but we can simplify it further for $k = 2$. By looking for a geometric-shaped solution to the steady-state occupancy distribution, we find:

$$p_{Q_{i,j}}(q_{i,j}) = (1 - \gamma_j) \gamma_j^{q_{i,j}}, \quad (40)$$

where γ_j is given by:

$$\gamma_j = \frac{2\lambda + \mu_j - \sqrt{\mu_j(\mu_j + 4\lambda)}}{2\mu_j}. \quad (41)$$

As the system time for a packet conditioned on the number of packets it finds in the queue is Erlang distributed, we can get the PDF of the system delay for a given queue:

$$p_{T^{\text{alt}},j}(t) = \sum_{q_{i,j}=0}^{\infty} (1 - \gamma_j) \frac{\mu_j (\mu_j \gamma_j t)^{q_{i,j}} e^{-\mu_j t}}{q_{i,j}!}. \quad (42)$$

Since the sum has a known solution, we can compute the overall PDF:

$$p_{T^{\text{alt}}}(t) = \frac{\sum_{j=1}^2 (1 - \varepsilon_j) \mu_j (1 - \gamma_j) e^{-\mu_j (1 - \gamma_j) t} u(t)}{\sum_{j=1}^2 (1 - \varepsilon_j)}. \quad (43)$$

In the alternating case, deriving even a bound for the PAoI is extremely complex, as even finding the state of the queues conditioned on y_i becomes intractable. Due to this, we do not provide analytical results for the PAoI in this case.

B. Split transmission

Synchronized $M/M/2$ systems follow the Flatto-Hahn-Wright (FHW) model, which is known to be intractable; Flatto [36], [37] computed the generating function of the steady-state distribution of the number of packets in the two queues, but there is no closed-form solution for either the state or the waiting time distributions.

We consider a *lower bound* on the PAoI by using the steady-state probability distribution of the equivalent $M/M/2/L$ blocking fork-join queue. Naturally, this is a lower bound, because we do not consider the blocked packets, but the bound is tight if the blocking probability is low, i.e., if $P(q_{i,j} > L) \ll 1$, where $q_{i,j}$ is the number of queued packets that packet i finds when it arrives to system j . Under this condition the system almost never reaches a full queue and operates like the infinite buffer case. The *upper bound* is obtained by considering all cases in which the queue reaches

L as a sample with infinite age. The same bounds are valid for both the $M/M/2/L$ and the $M/M/2/\infty$ fork-join system, and are tighter if the traffic load is low.

The $M/M/2/L$ blocking queue can be represented by a finite semi-Markov chain with state $\mathbf{Q} = (Q_1, Q_2)$, where Q_j represents the number of packets currently in system j , and transition matrix \mathbf{P} . We have three possible transitions, with all other elements of the transition matrix being equal to 0: either a packet arrives on both systems, or a packet leaves on one of the two systems. This relies on the common assumption of non-simultaneous departures, and on the independence of the two systems' service times. The computation of the probabilities in \mathbf{P} and the sojourn time in each state is trivial, and we will not report it here. We can then get the steady-state probability, which we denote as ϕ , as the solution to the equation $\phi(\mathbf{P} - \mathbf{I}) = 0$, normalized so that $\sum_{q_{i,1}=0}^L \sum_{q_{i,2}=0}^L \phi(\mathbf{q}_i) = 1$. This corresponds to the left eigenvector of \mathbf{P} with eigenvalue 1. The steady-state distribution π is obtained by weighting the distribution ϕ by the average sojourn times $\tau(\mathbf{q})$:

$$\pi(\mathbf{q}_i) = \frac{\phi(\mathbf{q}_i)\tau(\mathbf{q}_i)}{\sum_{\mathbf{q} \in \{0, \dots, L\}^2} \tau_{\mathbf{q}}}, \quad \forall \mathbf{q}_i \in \{0, \dots, L\}^2. \quad (44)$$

The PDF of the system time is simple, as if we condition on the state of the queue, we get the same PDF we derived in IV-B, as the arrival process does not affect the conditioned distribution. We first get the lower bound on the system time PDF:

$$p_{T_i}(t_i) = \sum_{q_{i,1}=0}^L \sum_{q_{i,2}=0}^L \pi(\mathbf{q}_i) p_{T_i|\mathbf{Q}_i}(t_i|\mathbf{q}_i). \quad (45)$$

The success probability is within $[(1 - \varepsilon_1 \varepsilon_2)(1 - \pi_{L,L}), 1 - \varepsilon_1 \varepsilon_2]$. The lower bound we derive for the PAoI considers an error-free system, as the errors complicate the derivation even further, but it holds for error-prone systems as well, although it is looser if we consider high packet error rates.

$$\begin{aligned} p_{\Delta_i|\mathbf{Q}_i}(\delta|(0,0)) &= \int_0^\delta p_{Y_i}(y) \sum_{q_{i-1,1}=0}^L u(\delta) \sum_{q_{i-1,2}=0}^L p_{T_i|\mathbf{Q}_i}(\delta - y|(0,0)) \\ &\quad \times \frac{\pi(\mathbf{q}_{i-1})}{\pi((0,0))} P_{\mathbf{Q}_i|\mathbf{Q}_{i-1}, Y_i}((0,0)|\mathbf{q}_{i-1}, y) dy \\ &= \sum_{\mathbf{q}_{i-1}=(0,0)}^{(L,L)} \frac{\lambda u(\delta) \pi(\mathbf{q}_{i-1})}{\pi((0,0))} \left[\sum_{j=1}^2 \left(\frac{\mu'_j (e^{-\lambda\delta} - e^{-\mu'_j\delta})}{\mu'_j - \lambda} \right. \right. \\ &\quad \left. \left. + \sum_{k=0}^{q_{i-1,j}} \frac{(\mu'_j)^k}{k!} \left((\mu'_1 + \mu'_2) e^{-(\mu'_1 + \mu'_2)\delta} \theta_{k,0}^{(\mu'_{3-j}-\lambda)}(\delta) \right. \right. \right. \\ &\quad \left. \left. - \mu'_j e^{-\mu'_j\delta} \theta_{k,0}^{(-\lambda)}(\delta) - \mu'_{3-j} e^{-\mu'_{3-j}\delta} \theta_{k,0}^{(\mu'_{3-j}-\mu'_j-\lambda)}(\delta) \right) \right) \\ &\quad \left. - \frac{(\mu'_1 + \mu'_2)(e^{-\lambda\delta} - e^{-(\mu'_1 + \mu'_2)\delta})}{\mu'_1 + \mu'_2 - \lambda} - \sum_{\mathbf{k}=(0,0)}^{\mathbf{q}_{i-1}} \prod_{j=1}^2 \left(\frac{(\mu'_j)^{k_j}}{k_j!} \right) \right. \\ &\quad \left. \left(\frac{\mu'_1 + \mu'_2}{e^{(\mu'_1 + \mu'_2)\delta}} \theta_{j+k,0}^{(-\lambda)}(\delta) - \sum_{j=1}^2 \frac{\mu'_j}{e^{\mu'_j\delta}} \theta_{j+k,0}^{(-\mu'_{3-j}-\lambda)}(\delta) \right) \right]. \end{aligned} \quad (46)$$

We can now consider the case in which only the second queue is occupied:

$$\begin{aligned} p_{\Delta_i|\mathbf{Q}_i}(\delta|(0, q_{i,2})) &= \sum_{\mathbf{q}_{i-1}=(0, q_{i,2}-1)}^{(L,L)} \lambda u(\delta) \frac{\pi(\mathbf{q}_{i-1})(\mu'_2)^{\psi_{i,2}}}{\pi(\mathbf{q}_i) \psi_{i,2}!} \left[\mu'_1 e^{-\mu'_1\delta} \right. \\ &\quad \times \left(\theta_{\psi_{i,2},0}^{(\mu'_1 - \mu'_2 - \lambda)}(\delta) - \sum_{m=0}^{q_{i-1,1}} \frac{(\mu'_1)^m}{m!} \theta_{\psi_{i,2}+m,0}^{(-\mu'_2 - \lambda)}(\delta) \right) \\ &\quad \left. + e^{-\mu'_2\delta} \left(\left(\theta_{\psi_{i,2},q_{i,2}}^{(-\lambda)}(\delta) - e^{-\mu'_1\delta} \theta_{\psi_{i,2},q_{i,2}}^{(\mu'_1 - \lambda)}(\delta) - \sum_{m=0}^{q_{i-1,1}} \frac{(\mu'_1)^m}{m!} \right. \right. \right. \\ &\quad \times \frac{(\mu'_2)^{q_{i,2}+1}}{q_{i,2}!} \left(\theta_{\psi_{i,2}+m,q_{i,2}}^{(-\mu'_1 - \lambda)}(\delta) - e^{-\mu'_1\delta} \theta_{\psi_{i,2}+m,q_{i,2}}^{(-\lambda)}(\delta) \right) \right) \\ &\quad \left. \left. - \sum_{n=0}^{q_{i,2}} \frac{\mu'_1 (\mu'_2)^n}{e^{\mu'_1\delta} n!} \left(\theta_{\psi_{i,2},n}^{(-\mu'_1 - \lambda)}(\delta) - \sum_{m=0}^{q_{i-1,1}} \frac{\theta_{\psi_{i,2}+m,n}^{(-\lambda)}(\delta)}{(\mu'_1)^{-m} m!} \right) \right) \right]. \end{aligned} \quad (47)$$

If only the first queue is occupied, the conditioned PDF is the same, with switched indices. Finally, we consider the case in which both queues are occupied:

$$\begin{aligned} p_{\Delta_i|\mathbf{Q}_i}(\delta|\mathbf{q}_i) &= \sum_{\mathbf{q}_{i-1}=\mathbf{q}_i-1}^{(L,L)} \lambda u(\delta) \sum_{j=1}^2 \frac{(\mu'_j)^{\psi_{i,j}} (\mu'_{3-j})^{q_{i,3-j}+1}}{\psi_{i,1}! \psi_{i,2}! e^{\mu'_{3-j}\delta} q_{i,3-j}!} \\ &\quad \times \left[\theta_{\psi_{i,1},q_{i,3-j}}^{(-\mu'_j - \lambda)}(\delta) - \sum_{n=0}^{q_{i,j}} \frac{\theta_{\psi_{i,1},q_{i,3-j}+n}^{(-\lambda)}(\delta)}{(\mu'_j)^{-n} e^{\mu'_j\delta} n!} \right] \frac{\pi(\mathbf{q}_{i-1})}{\pi(\mathbf{q}_i)}. \end{aligned} \quad (48)$$

The overall age bound can be computed simply by applying the law of total probability:

$$p_{\Delta_i}(\delta) = \sum_{q_{i,1}=0}^L \sum_{q_{i,2}=0}^L p_{\Delta_i|\mathbf{Q}_i}(\delta|\mathbf{q}_i) \pi(\mathbf{q}_i). \quad (49)$$

C. Replicated transmission

As for the split scheme, the arrival process does not affect the conditioned distribution of the system time with a given queue state. The lower bound on the system time PDF then follows the same form as in (45), using the conditioned PDF derived in Sec. IV-C. In this case, the success probability is within $[(1 - \varepsilon_1 \varepsilon_2)(1 - \pi_{L,L}), 1 - \varepsilon_1 \varepsilon_2]$. As for the split case, we compute a lower bound for the PAoI which does not take errors into account and consider the four cases that we used above to derive the PAoI lower bound, starting from the one with an empty queue:

$$\begin{aligned} p_{\Delta_i|\mathbf{Q}_i}(\delta|(0,0)) &= u(\delta) \sum_{q_{i-1,1}=0}^L \frac{(\mu'_1 + \mu'_2) \lambda e^{-(\mu'_1 + \mu'_2)\delta}}{\pi(\mathbf{q}_i)} \\ &\quad \times \sum_{q_{i-1,2}=0}^L \pi(\mathbf{q}_{i-1}) \left[\frac{(e^{(\mu'_1 + \mu'_2 - \lambda)\delta} - 1)}{\mu'_1 + \mu'_2 - \lambda} - \sum_{k=0}^{q_{i-1,1}} \frac{(\mu'_1)^k}{k!} \theta_{k,0}^{(\alpha_2)}(\delta) \right. \\ &\quad \left. - \sum_{k=0}^{q_{i-1,2}} \frac{(\mu'_2)^k}{k!} \left(\theta_{k,0}^{(\alpha_1)}(\delta) - \sum_{j=0}^{q_{i-1,1}} \frac{(\mu'_1)^j \theta_{j+k,0}^{(-\lambda)}(\delta)}{j!} \right) \right]. \end{aligned} \quad (50)$$

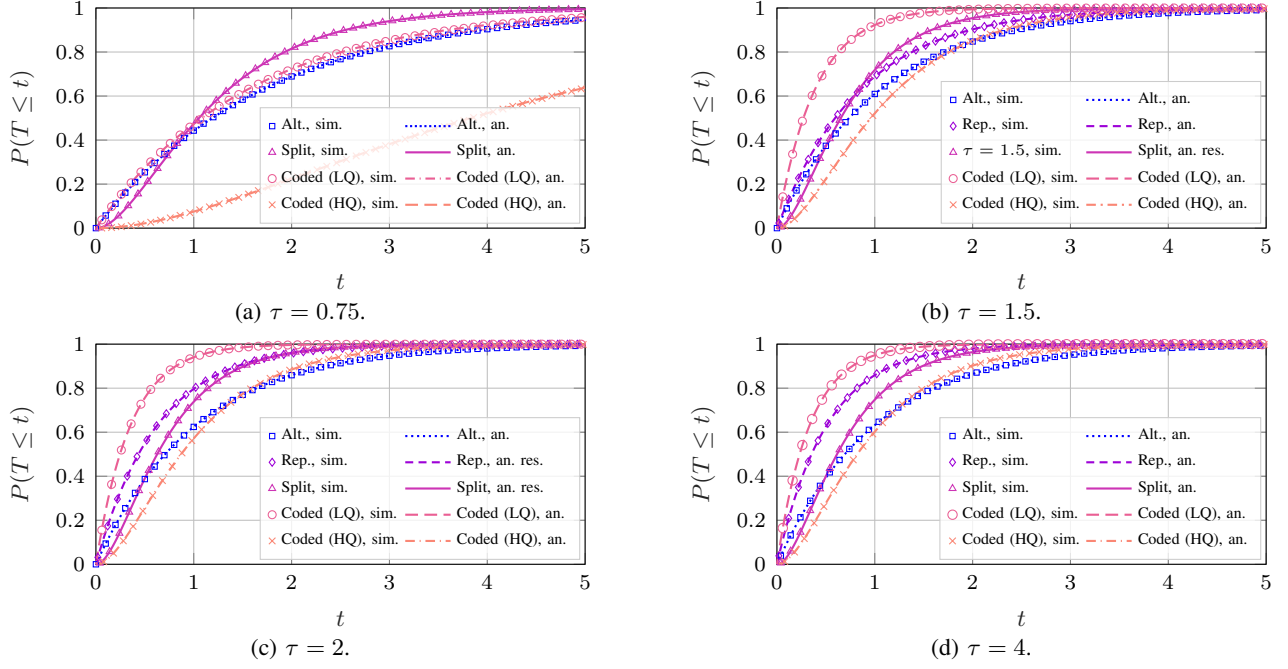


Fig. 2: Latency CDF in an error-free $D/M/2$ with $\mu_1 = \mu_2 = 1$.

We then look at the case in which only the first queue is occupied:

$$p_{\Delta_i | \mathbf{Q}_i}(\delta | (q_{i,1}, 0)) = \sum_{q_{i-1,1}=q_{i,1}-1}^L u(\delta) \lambda \sum_{q_{i-1,2}=0}^L \frac{\pi(\mathbf{q}_{i-1}) e^{-(\mu'_1 + \mu'_2)\delta}}{\pi(\mathbf{q}_i) \psi_{i,1}!} \left[\sum_{j=0}^{q_{i,1}} \frac{\mu'_1(\mu'_2)^{\psi_{i,1}+j}}{j!} \left(\theta_{\psi_{i,1},j}^{(\alpha_2)}(\delta) - \sum_{k=0}^{q_{i-1,2}} \frac{(\mu'_2)^k \theta_{\psi_{i,1}+k,j}^{(-\lambda)}(\delta)}{k!} \right) + \frac{(\mu'_1)^{q_{i-1,1}+2}}{q_{i,1}!} \left(\theta_{\psi_{i,1},q_{i,1}}^{(\alpha_2)}(\delta) - \sum_{k=0}^{q_{i-1,2}} \frac{(\mu'_2)^k \theta_{\psi_{i,1}+k,q_{i,1}}^{(-\lambda)}(\delta)}{k!} \right) \right]. \quad (51)$$

Naturally, the case in which only the second queue is occupied is given equivalently, with inverted indices. Finally, if both queues are occupied, the conditioned bound PDF is:

$$p_{\Delta_i | \mathbf{Q}_i}(\delta | (q_{i,1}, q_{i,2})) = \lambda \sum_{\mathbf{q}_{i-1}=\mathbf{q}_i-1}^L u(\delta) \left(\prod_{j=1}^2 \frac{(\mu'_j)^{q_{i-1,j}+2}}{e^{\mu'_j \delta} \psi_{i,j}!} \right) \times \frac{\pi(\mathbf{q}_{i-1})}{\pi(\mathbf{q}_i)} \sum_{j=1}^2 \sum_{k=0}^{q_{i,j}} \frac{\theta_{\psi_{i,j},q_{i,3-j}+k}^{(-\lambda)}(\delta)}{(\mu'_j)^{q_{i,j}+1-k} q_{i,3-j}! k!}. \quad (52)$$

The full unconditioned PDF of the lower bound is then derived by applying the law of total probability over the queue states, as in (49).

D. Coded transmission

As for the $D/M/2$ case, the coded transmission is simply a combination of the previous two cases, and we can skip the derivation of the latency and PAoI for the low- and high-quality versions. The procedure to derive the analytical bounds is equivalent to what we explained in the above for the replicated and split schemes, adjusted with the appropriate value of η .

VI. NUMERICAL RESULTS

In this section, we compare our analytical results with an extensive Monte Carlo simulation, consisting of over 10^6 packets, in order to verify the analysis. The first 1000 packets of each simulation were cut from the results to avoid the initial transition effects and ensure that the system was only considered in a steady state. We analyzed both the latency and the PAoI for the four schemes, considering both the full distribution and the 99th percentile, used as a proxy for the worst-case performance. The value of L that we used for the $M/M/2$ approximated bounds was 10: higher values would provide a tighter bound for higher loads, but also significantly increase the computational complexity of the bound calculation. Unless otherwise stated, the coded transmission scheme uses $\eta = 0.75$ in all plots. For the cases in which we have derived bounds and not the exact values in Sec. IV and Sec. V, we have compared the bounds to the Monte Carlo simulation results, showing that the bounds are tight in the regions of interest (i.e., when the system is not extremely congested), and are therefore useful for system design purposes. Due to space limitations, we do not show the average AoI results, but we have verified them by simulation, and the same considerations that we make in the following on the tightness of the PAoI bounds hold. We also remind the reader that low- and high-quality latency and age (LQ and HQ, respectively) are considered separately in the results.

A. Latency

We first look at the latency of the four schemes in $D/M/2$ systems. Fig. 2 shows the CDF of the system time for different values of the inter-arrival period τ in an error-free, symmetrical system. As expected, the empirical CDF of the system time perfectly fits the analytical derivation for all cases. If we look at the case with $\tau = 0.75$, shown in Fig. 2a, the split scheme is

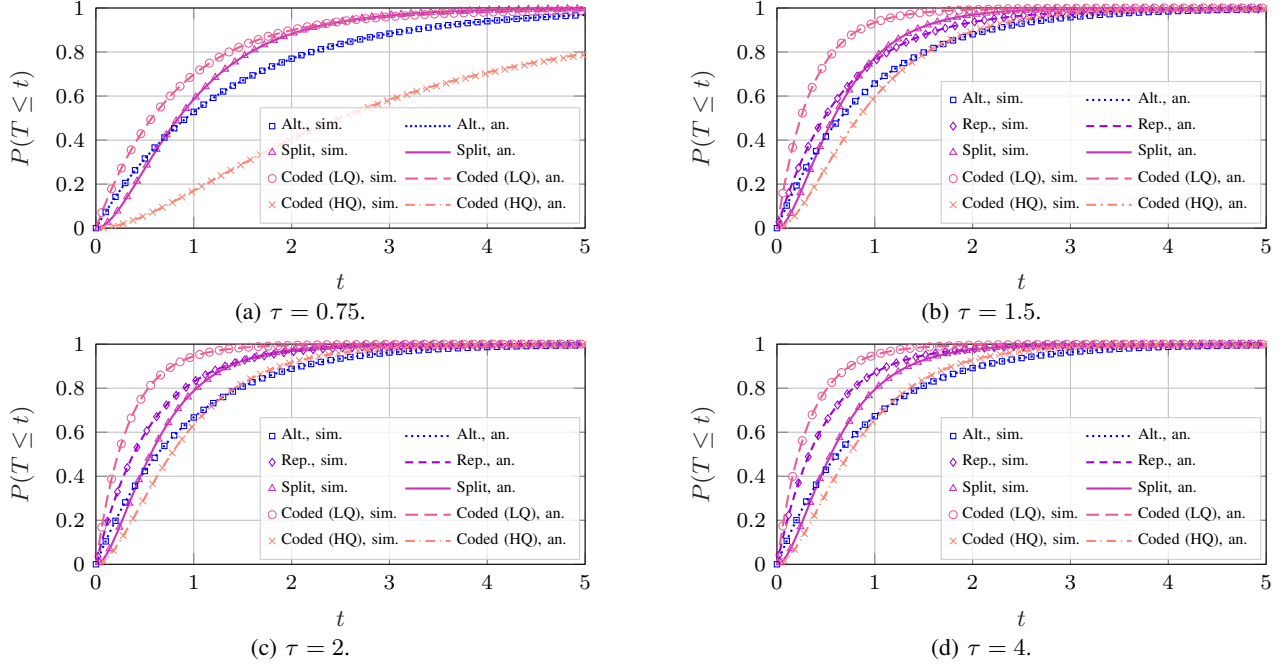


Fig. 3: Latency CDF in an error-prone $D/M/2$ with $\mu_1 = 1$, $\mu_2 = 1.25$, and $\varepsilon_1 = \varepsilon_2 = 0.05$.

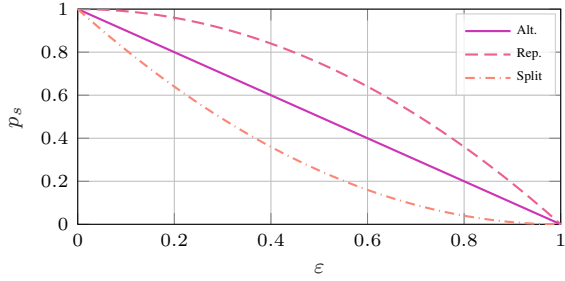


Fig. 4: Block delivery probability as a function of ε in an error-prone system with $\varepsilon_1 = \varepsilon_2 = \varepsilon$.

by far the best to provide reliable low latency: in this case, the queue is congested, and reducing the load is critical. Shorter packets can also reach the receiver faster, even if both need to be delivered: for quantiles over 0.4, the split scheme is faster than even the low-quality coded scheme, which only requires one packet of size $\eta = 0.75$ to reach the receiver. As the value of τ grows, from $\tau = 1.5$ in Fig. 2b to $\tau = 2$ in Fig. 2c and $\tau = 4$ in Fig. 2d, the replicated scheme, which was not even shown in the first plot because it is unstable for $\tau = 0.75$, gradually becomes a better option. This kind of strategy can deal with delays on one path by using the other replica as a backup, but it doubles the load on both paths, so it is only efficient if the inter-block period is long. We can see that the low-quality version of the coded scheme is always faster, as it sends smaller packets. On the other hand, the latency performance of the high-quality version is comparable to the alternating scheme, and far worse than either the split or replicated schemes for any value of τ over 1.5.

In Fig. 3, we look at what happens if there is an imbalance between the two systems, i.e., in a system with $\mu_1 = 1$ and $\mu_2 = 1.25$. We also introduce an error probability $\varepsilon = 0.05$ on both paths. In this case, the latency distribution presents some differences: the low-quality version of the coded scheme is

actually faster than the split scheme in the case with $\tau = 0.75$, shown in Fig. 3a. This is due to the higher rate of the second path, which can make the arrival of packets on both paths slower even if the two are smaller. If we consider larger values of τ , the considerations we can draw are similar to the previous case.

However, there is one major difference between the two systems: the CDFs shown above only consider successfully delivered blocks, and the schemes have highly different delivery probabilities, as Fig. 4 shows: as expected, the replicated scheme has a far higher delivery probability (and, consequently, so does the coded scheme if we consider the low-quality version of the block), while the split scheme has the worst performance, as it requires both packets to be delivered to decode the block. The coded scheme has the same performance as the split scheme if we only consider the high-quality version of the blocks.

We can now look at latency in an $M/M/2$ fork-join system: as Fig. 5 shows, the basic trade-offs between the schemes are different, and the tail of the latency distribution is longer: this is expected, as randomness in the arrivals can negatively affect the worst-case latency. In this case, the analytical curves represent a lower bound to the latency. However, the bound is extremely tight for $\tau \geq 1.5$, and all four schemes show a negligible difference with the Monte Carlo results in those cases. The bound is not tight if we consider the coded scheme with $\tau = 0.75$, as Fig. 5a shows: in this case, the queues can become congested, as the load ρ_j on each link is 0.89, and the bound is extremely optimistic. Note that any case with ρ_j close to 1 is highly suboptimal for both latency and PAoI, as the system is close to congestion. We conclude that the bounds are useful in the regions of interest for performance optimization, when the queues are not congested, and become loose as ρ_j approaches 1. Interestingly, the performance considerations are different in this case, as the alternating and split scheme

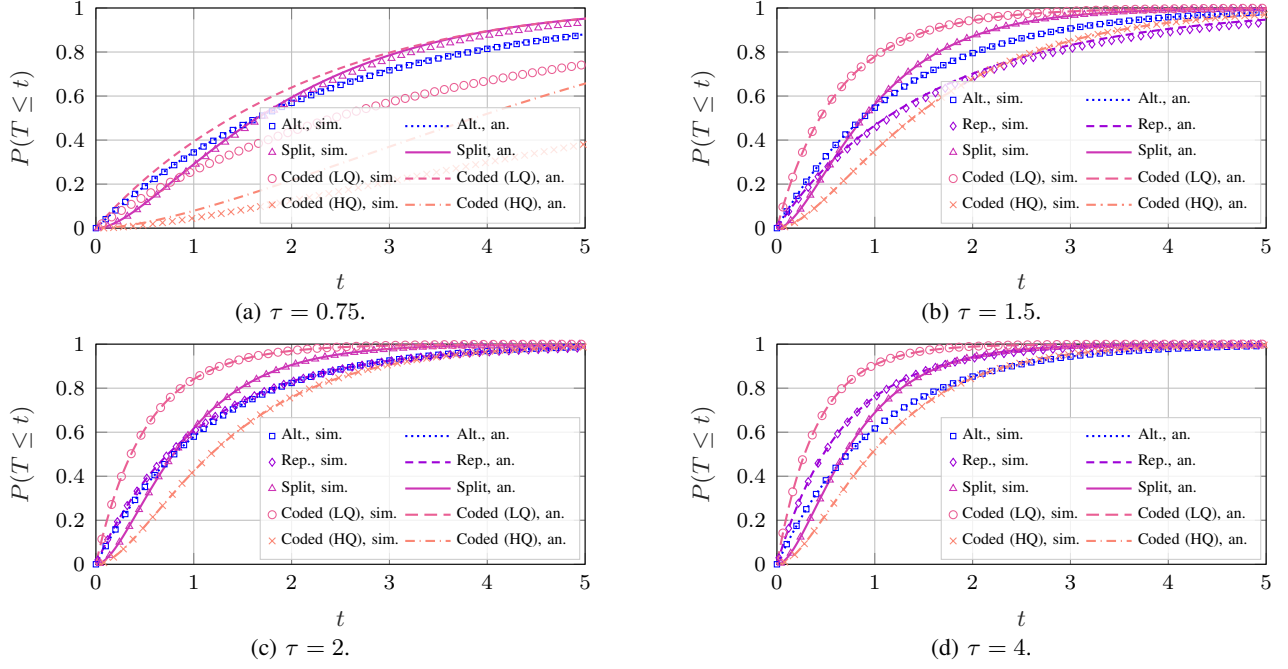


Fig. 5: Latency CDF in an error-free $M/M/2$ with $\mu_1 = \mu_2 = 1$ and $\tau = \frac{1}{\lambda}$.

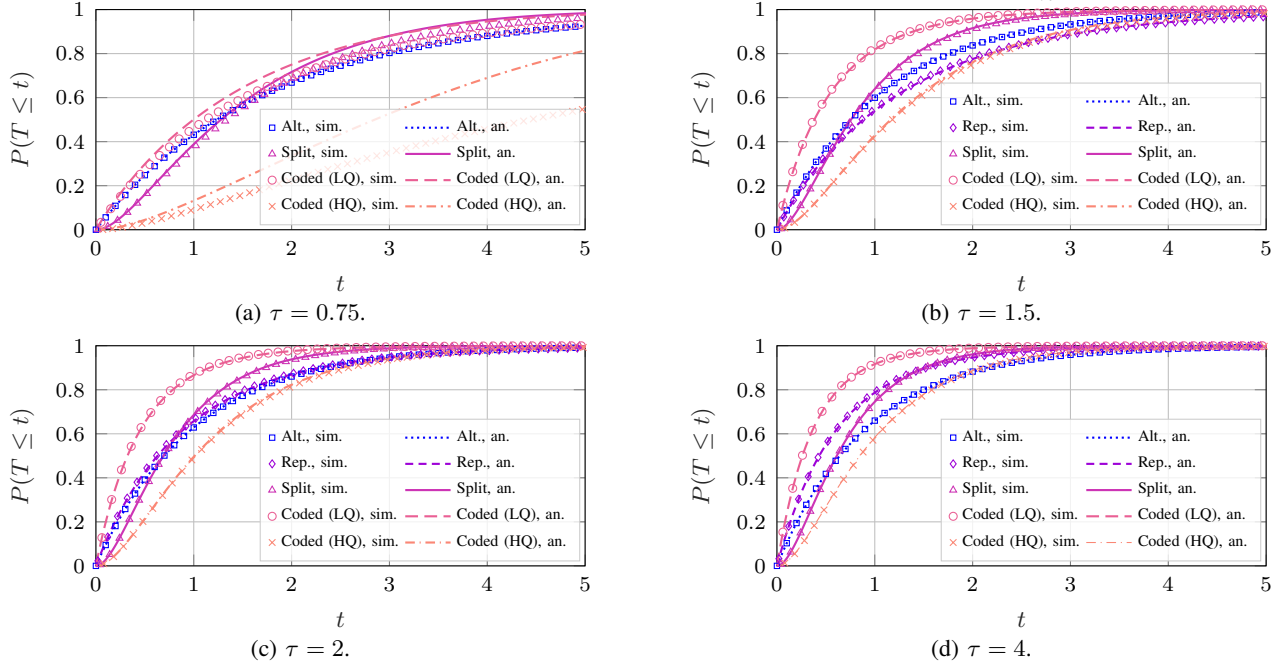


Fig. 6: Latency CDF in an error-prone $M/M/2$ with $\mu_1 = 1$, $\mu_2 = 1.25$, and $\varepsilon_1 = \varepsilon_2 = 0.05$.

perform better with respect to the coded and replicated ones: as Markovian arrivals can lead to several arrivals in a short time, reducing the load on the system is more important, giving the split and alternating schemes an advantage.

If we consider the unbalanced system with $\mu_2 = 1.25$ and $\varepsilon = 0.05$, we can see from Fig. 6 that the improvement in performance due to the higher service rate of the second system is much higher. However, the relative advantage of the alternating and split schemes is smaller, and the tail of the alternating scheme distribution is comparatively longer, as one of the two paths is now faster, and those schemes require packets on both paths to arrive. On the other hand, the

replicated scheme can fully exploit the faster path, reducing the disadvantage from the higher load it generates. The latency bounds are still tight, aside from the case with $\tau = 0.75$, and can be used as practical system design guidelines with a negligible approximation, as the best configurations for latency have a relatively low traffic rate λ , and the tightness of the bound increases as the load on the system decreases.

Naturally, the effectiveness of the coded system depends on the efficiency of the code: a coding scheme with a higher η will have smaller packets, but each individual packet will contain less information on the block, so the lower-quality version will have a worse quality than the equivalent for a

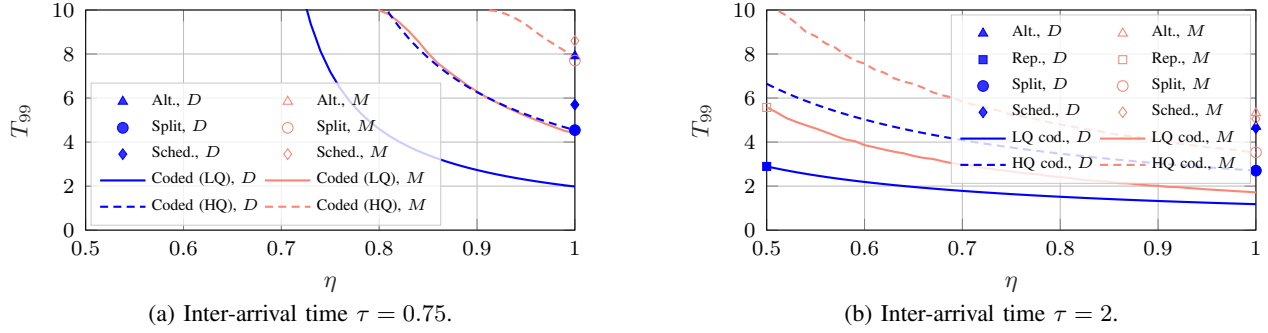


Fig. 7: 99th percentile T_{99} of the latency as a function of the coding rate η in an error-free system with $\mu_1 = \mu_2 = 1$. The $D/M/2$ and $M/M/2$ systems are plotted using different colors.

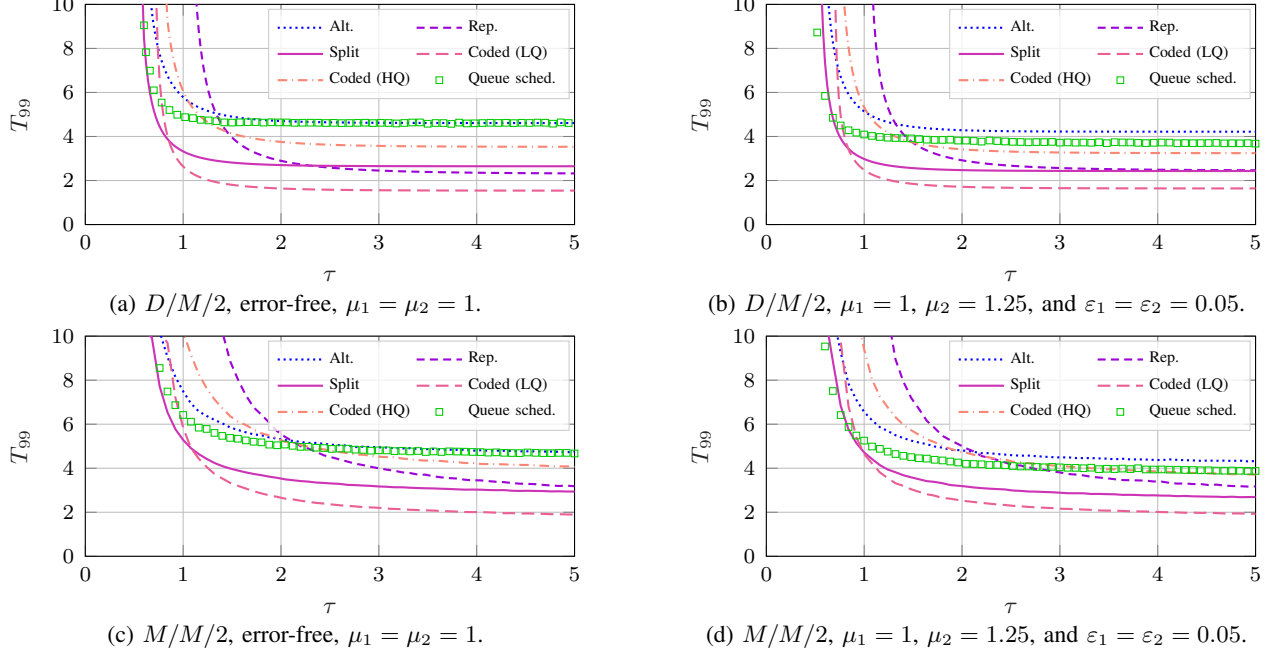


Fig. 8: 99th percentile T_{99} of the latency as a function of the inter-block time in $D/M/2$ and $M/M/2$ systems.

scheme with a smaller η . The replicated and split scheme are the two extremes: in the replicated scheme, each packet is enough to get the high-quality version of the block on its own, while in the split scheme, an individual packet is not enough to decode the block at any quality. Fig. 7 shows the worst-case latency performance of the schemes, represented using the 99th percentile of the system time T_{99} for both deterministic and Markovian arrivals. We have $\eta = 1$ for the alternating and split schemes, and $\eta = 0.5$ for the replicated scheme. As the plots show, there is an inverse relationship between the latency performance and the quality achievable with just one packet, as increasing the redundancy leads to a sharp increase in the latency, and the system can become unstable if packets are both frequent and large, as is shown in Fig. 7a, which has $\tau = 0.75$. In this case, the load on the two paths is already high, so choosing a low η can increase the queuing delay significantly, with a significant delay increase. If the two paths have a very low load, as in Fig. 7b, which has $\tau = 2$, the benefits from a more efficient code are smaller, and the latency cost of having a higher quality from a single packet is lower. As we noted above, the $M/M/2$ system is not ideal if the goal is worst-case performance: as the uncertainty on arrival times

is higher, the high percentiles of the latency suffer significantly with respect to the $D/M/2$ case.

We can also look at system optimization, if we have an application that can work at different arrival rates: Fig. 8 shows the value of T_{99} as a function of τ for the $D/M/2$ and $M/M/2$ systems. As expected, a longer inter-arrival time leads to a reduced latency in all cases: as the load on the system decreases, the probability of having to wait for the queued blocks to be sent becomes smaller. The alternating system, which does not exploit the parallel paths fully, can support a high update frequency but performs worse than the other systems when the load is low (the 99th percentile of the service time, with $\mu = 1$, would be equal to $-\log(0.01) \simeq 4.6$).

On the other hand, the replicated system has a very high load, and underperforms for low values of τ , but can actually keep a low latency when the load on the system is low. The split system provides a middle ground, but it is also highly vulnerable to errors, as previously discussed. Finally, the coded solution can provide a very good delay, smaller than all the other schemes', for the low-quality version, while still delivering the high-quality block in a reasonable time. We also note that these plots only consider successfully delivered

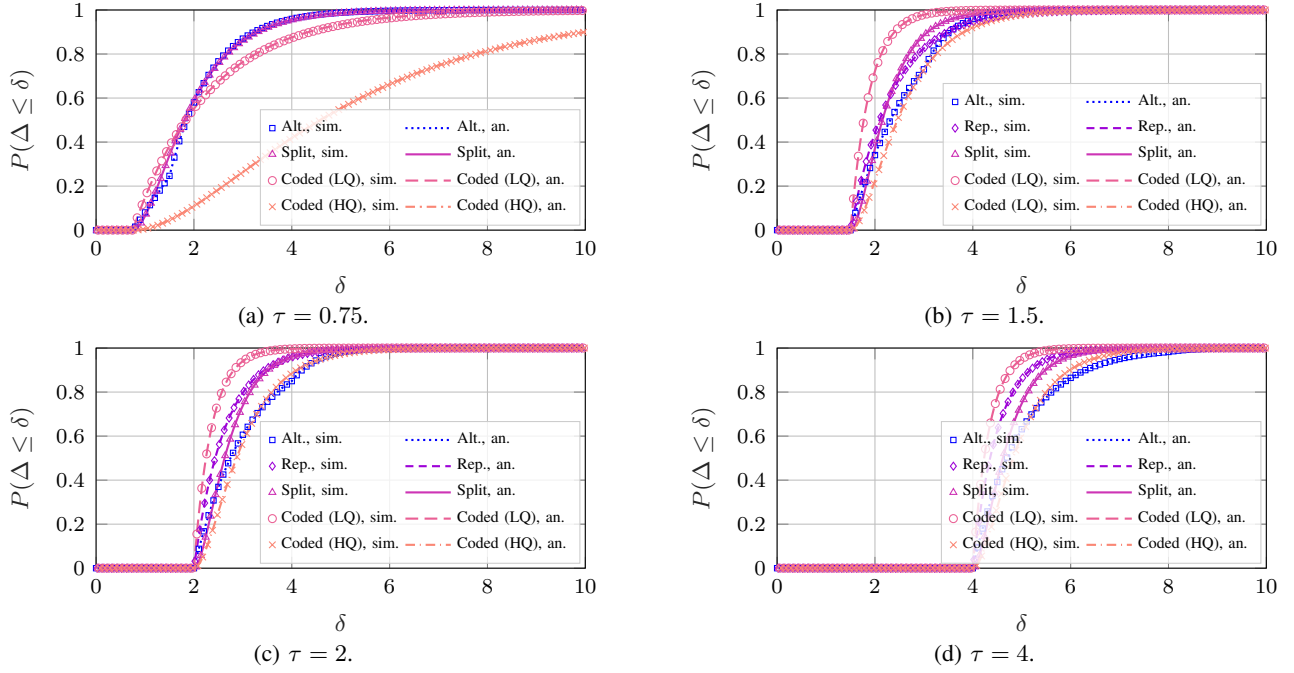


Fig. 9: PAoI distribution in an error-free $D/M/2$ with $\mu_1 = \mu_2 = 1$.

blocks, and need to be weighted by the delivery probability from Fig. 4, so that the robustness of the replicated scheme makes it even more attractive to minimize worst-case latency in error-prone scenarios. The same considerations apply for both the $D/M/2$ and $M/M/2$ system, but performance in the high-load scenarios is significantly worse, due to the higher uncertainty in the arrival process, which can lead to more significant queue buildups. This can also be observed by comparing the performance of a shortest-queue scheduler (denoted as *Queue sched.* in the figures), which always picks the path with the shortest queue, with the pure round-robin scheduler like the alternating scheme: while the alternating scheduler works very well in $D/M/2$ systems, except for very high loads, the difference in an $M/M/2$ is significant, although the coded and split scheme can still outperform uncoded transmission in almost all settings.

B. Peak Age of Information

We can now look at PAoI, analyzing the performance of the four schemes for the $D/M/2$ and $M/M/2$ systems. We first look at the simplest case of an error-free $D/M/2$, in which the two systems have the same service rate. Fig. 9 shows the empirical and analytical PAoI CDFs in this scenario, in which it is possible to compute the PAoI distribution exactly only for the replicated scheme. However, the probability of multiple out-of-order packets is very low, and the bound is extremely tight even for this case. We can easily see that the considerations we can draw are highly different than for the latency: the alternating scheme becomes a very good choice, particularly when the load on the system is high, as late or out-of-order packets can be superseded by the other path. On the other hand, high values of τ correspond to a far worse PAoI performance, as the inter-arrival time becomes the dominating component of the age, even as latency decreases. It

is interesting to note that the CDF for the alternating scheme is not smooth: this is because in some cases block i arrives before block $i-1$, which is transmitted on the other path. This out-of-order arrival can only result in a PAoI of at least 2τ , as the age of the $i-2$ -th block, which has certainly arrived before the i -th as it was sent on the same path, is 2τ when the i -th block is sent. As the figure shows, aside from the low-quality version of the coded scheme, which outperforms the other schemes in most cases, the split and alternating schemes perform best for $\tau = 0.75$ and $\tau = 1.5$, while the replicated scheme is slightly better for longer inter-block periods. Average and worst-case performance are often very different, and optimizing for the average AoI or PAoI will not necessarily result in good worst-case performance.

If we consider the system with $\mu_2 = 1.25$ and $\varepsilon = 0.05$, shown in Fig. 10, we can notice that the split and alternating scheme perform worse, as they need to rely on the slower path and are much more affected by errors: in this case, the replicated scheme works better. Finally, the CDF for the split scheme also loses its smoothness in this case: if a block is lost due to one of the packets being erased by the channel, the system needs to wait another interval τ before the next update, causing a non-differentiable point at each multiple of τ in the peak age CDF. In both of these systems, the PAoI almost perfectly matches the bounds, indicating that the bounds are tight and can be used for system design purposes.

In the $M/M/2$ system, age is significantly higher: we can see this for the error-free balanced system in Fig. 11a-b. We do not show the cases with $\tau = 2$ and $\tau = 4$, as the age in those cases is extremely high. In this case, the analytical curves are all bounds, but the difference between the bounds and the analytical results is negligible, except for the high-quality version in the coded scheme with $\tau = 0.75$, for the same reasons we explained above when discussing the latency.

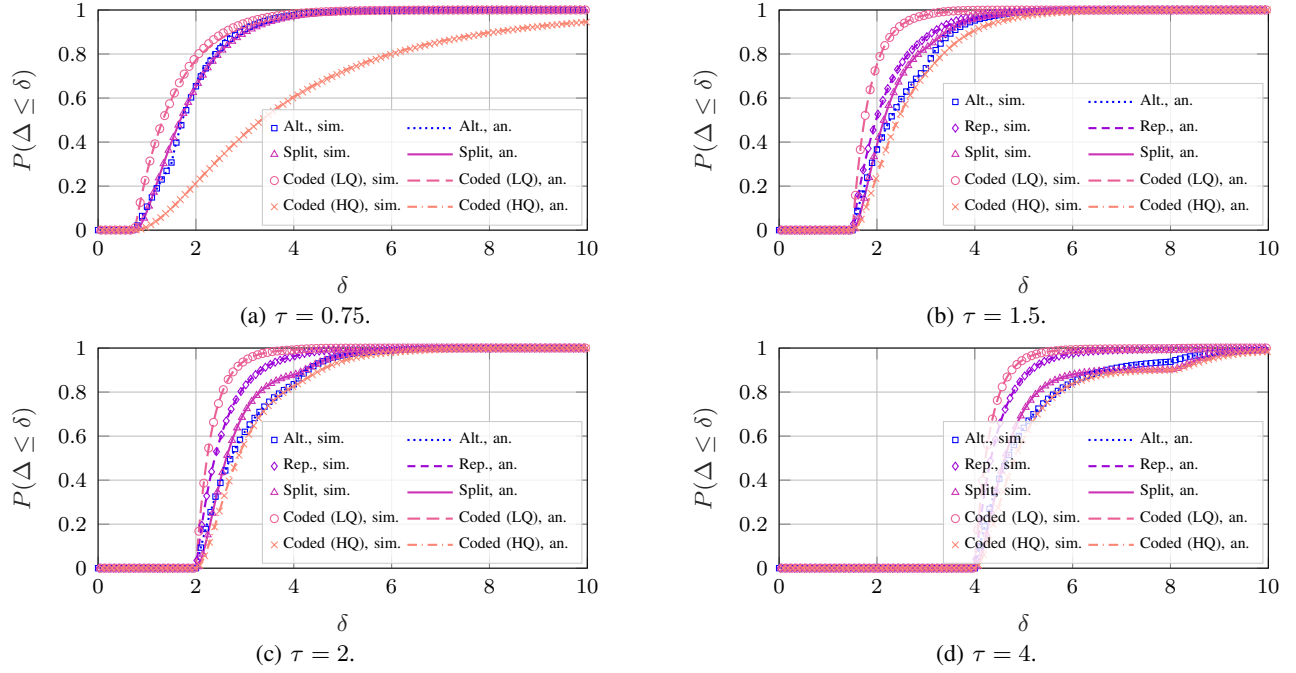


Fig. 10: PAoI distribution in an error-prone $D/M/2$ with $\mu_1 = 1$, $\mu_2 = 1.25$, and $\varepsilon_1 = \varepsilon_2 = 0.05$.

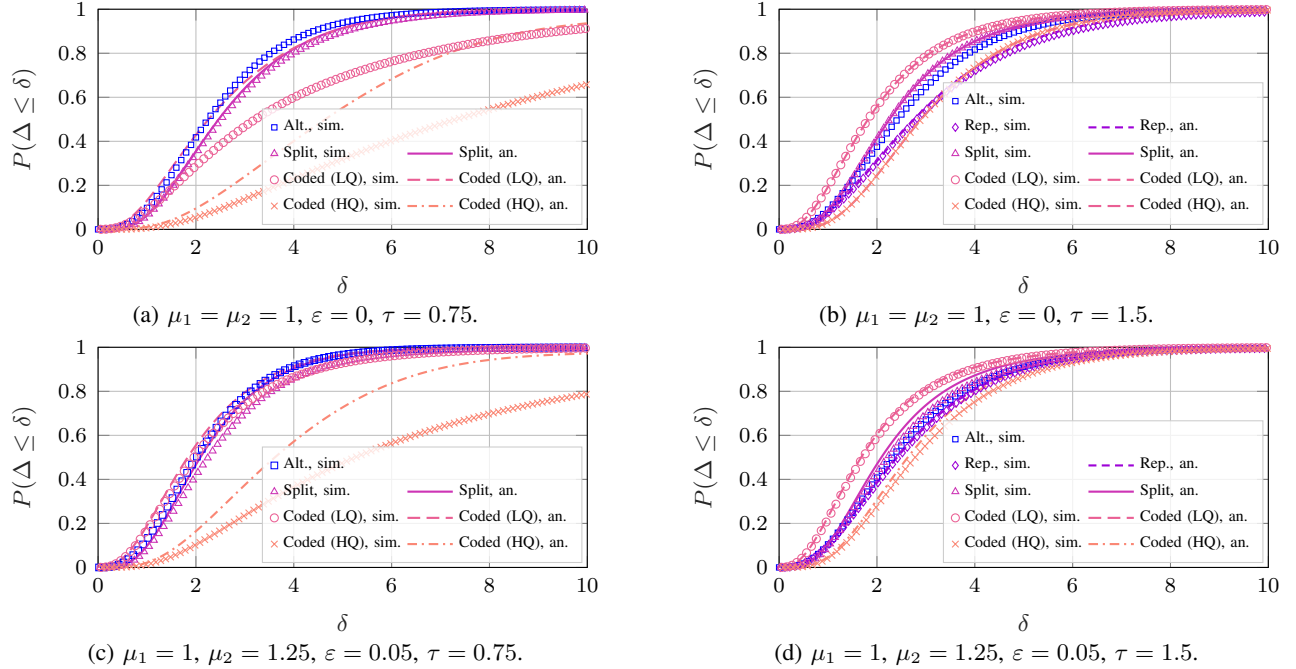


Fig. 11: PAoI distribution in an $M/M/2$ system.

We can also notice that the replicated scheme underperforms, as a queuing delay on either system can cause the age to increase significantly, and maintaining a low load is extremely important to reduce the age. The same considerations can be drawn for the error-prone unbalanced system in Fig. 11c-d, although the bounds are looser in this case: as the bounds only consider an error-free system, this is noticeable for the split scheme, which has the highest block failure probability. However, the bounds are still reasonably tight, except for the coded scheme with $\tau = 0.75$. In this case, the split, alternating, and replicated scheme all seem to have similar performance for $\tau = 1.5$, while the alternating scheme is slightly better

than the split for $\tau = 0.75$.

We can now look at the optimization of the system, using the 99th percentile of the PAoI Δ_{99} as a metric for the worst-case performance. Fig. 12 shows what happens as we change τ for the four schemes, along with the scheduled scheme. All results have a U shape, with an optimal block frequency which balances the inter-block period and the latency of blocks that are sent. Naturally, this optimal point is different for different schemes and scenarios. However, it is interesting to note that the lower-quality version of the coded scheme always has the lowest age, aside from having the lowest block loss. On the other hand, the replicated scheme always performs

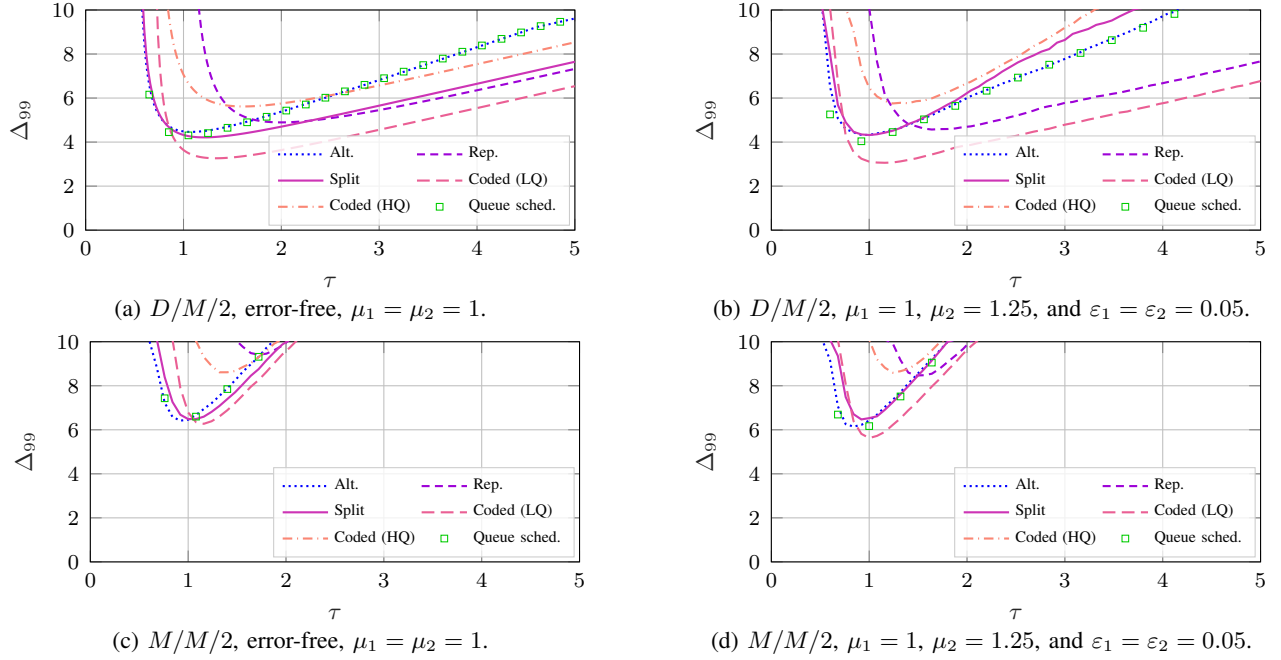


Fig. 12: 99th percentile Δ_{99} of the PAoI as a function of the inter-arrival time in the $D/M/2$ and $M/M/2$ systems. The coded transmission has $\eta = 0.75$.

slightly worse than the split scheme, as even though it will lose fewer blocks, its higher load significantly increases the optimal inter-block period. We can also see that the real age for the alternating system is close to the bound for the plots with error (shown on the right). The alternating scheme can perform slightly better than the split scheme, but is outperformed by the coded scheme at the lower quality. However, the high-quality version of the coded blocks has a very high age, as it has no protection from errors and the higher load associated to the redundancy in the coded scheme. In general, the alternating scheme performs surprisingly well in terms of PAoI. These trends hold for both the $D/M/2$ and $M/M/2$ system, but the latter has a significantly higher Δ_{99} , as we remarked above.

As we did for the system time, we can finally analyze the effect of the coding rate η on Δ_{99} in the coded scheme, as shown in Fig. 13. In this case, we have used an optimized arrival rate, i.e., set the inter-arrival time that minimizes the 99th percentile of the PAoI for each value of η . The plot on the left, which shows the error-free, balanced scenario, shows a smaller effect of the coding rate, which becomes more important for error-prone, unbalanced system. First, we note that, as for all other cases, the $M/M/2$ system has a far higher Δ_{99} than the $D/M/2$ one, due to the increased unpredictability in its behavior. The intuitive understanding that we got from the system time results holds: the harder the system conditions get, the more an efficient code matters, and the tighter the trade-off between quality and latency or age becomes. However, while latency-oriented systems are very sensitive to load, PAoI is very sensitive to the packet erasure probability, as missing a block can significantly increase the age. The results for the queue-based scheduler are also shown: in this case, the alternating scheme performs almost as well even in the unbalanced scenario. This is because sending multiple consecutive packets over the same path might improve

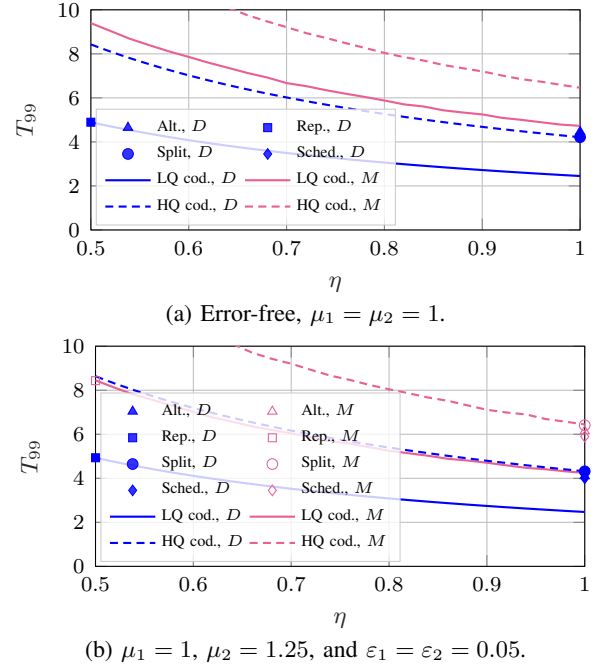


Fig. 13: 99th percentile Δ_{99} of the PAoI with optimized block frequency as a function of the coding rate η . The $D/M/2$ and $M/M/2$ systems are plotted with different colors.

the average age, but has a negative effect in the worst case, as one hold-up can block two consecutive packets, increasing the age significantly. The coded scheme could benefit from using a more intelligent scheduler that regulates η , reducing the redundancy if the queues are already congested: unlike latency-oriented systems, which should have a very low load (and, consequently, little effect from improving the scheduler, as the queues are almost always empty), PAoI is minimized for rather high load values, close to 0.5, so having an intelligent

scheduler can be a significant improvement.

VII. CONCLUSIONS AND FUTURE WORK

In this work, we analyzed some schemes for the transmission of data over parallel queuing systems, with multipath communications as a motivating scenario. Unlike previous works on the fork-join model, we derived bounds to the full distribution of the latency and PAoI for uncoded and coded schemes in the presence of communication errors, and examined the performance of the various schemes as a function of the block frequency and the efficiency of the encoding. The trade-offs between data quality, inter-generation times, and timeliness are complex, and our analysis provides a handy tool for system designers.

In our results, we show that the split policy can perform best in terms of both latency and PAoI, at the cost of reduced reliability in case of error-prone paths. On the other hand, selecting a single path, either by alternating between paths or choosing the shortest-queue path, can provide a relatively low PAoI at the cost of a higher latency. A coded scheme with a coding rate $\eta = 0.75$ can outperform all other schemes, at the cost of accepting some data distortion, but has a far higher latency and PAoI for the full-quality version due to the redundancy in the coding.

The analysis opens several avenues of future research, most of which aim at making the model more realistic. The use of more complex communication models and scheduling schemes is one, while another is a closer examination of real IoT and VR traffic models. These two extensions of the work can be combined to reach a realistic scheduling framework, which should also take into account the human perception of smoothness and delay in VR, and the effectiveness of control in IoT systems.

REFERENCES

- [1] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *International Conference on Computer Communications (INFOCOM)*. IEEE, Mar. 2012, pp. 2731–2735.
- [2] T. Chakraborti, S. Sreedharan, A. Kulkarni, and S. Kambhampati, "Projection-aware task planning and execution for human-in-the-loop operation of robots in a mixed-reality workspace," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2018, pp. 4476–4482.
- [3] A. Bröring, V. Kulkarni, A. Zirkler, K. Fysarakis, S. Mayer, B. Soret, L. D. Nguyen, P. Popovski, S. Samarakoon, M. Bennis, J. Harri, M. Rooker, G. Fritz, A. Bucur, G. Spanoudakis, and S. Ionannidis, "IntellIoT: Intelligent, real-time, and trusted IoT environments with human-in-the-loop," *Submitted to European Conference on Networks and Communications*, 2021.
- [4] M. Li, K. Arning, L. Vervier, M. Zieffle, and L. Kobbelt, "Influence of temporal delay and display update rate in an augmented reality application scenario," in *14th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 2015, pp. 278–286.
- [5] J. Wu, B. Cheng, C. Yuen, N.-M. Cheung, and J. Chen, "Trading delay for distortion in one-way video communication over the internet," *IEEE Trans. on Circuits and Sys. for Video Tech.*, vol. 26, no. 4, pp. 711–723, 2016.
- [6] F. Chiariotti, S. Kucera, A. Zanella, and H. Claussen, "Analysis and design of a latency control protocol for multi-path data delivery with pre-defined QoS guarantees," *IEEE/ACM Trans. on Net.*, vol. 27, no. 3, pp. 1165–1178, Jun. 2019.
- [7] F. Chiariotti, B. Soret, and P. Popovski, "Peak Age of Information distribution bounds for multi-connectivity transmissions," in *22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Sep. 2021, pp. 321–325.
- [8] M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, "On the role of age of information in the Internet of Things," *IEEE Communications Magazine*, vol. 57, no. 12, pp. 72–77, Dec. 2019.
- [9] Y. Inoue, H. Masuyama, T. Takine, and T. Tanaka, "A general formula for the stationary distribution of the Age of Information and its application to single-server queues," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 8305–8324, Dec. 2019.
- [10] A. Kosta, N. Pappas, V. Angelakis *et al.*, "Age of information: A new concept, metric, and tool," *Foundations and Trends in Networking*, vol. 12, no. 3, pp. 162–259, Nov. 2017.
- [11] R. D. Yates, "The age of information in networks: Moments, distributions, and sampling," *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5712–5728, May 2020.
- [12] X. Chen, K. Gatsis, H. Hassani, and S. S. Bidokhti, "Age of information in random access channels," in *International Symposium on Information Theory (ISIT)*. IEEE, Jun. 2020, pp. 1770–1775.
- [13] A. Munari, "Modern random access: an age of information perspective on Irregular Repetition Slotted ALOHA," *IEEE Transactions on Communications*, vol. 69, no. 6, pp. 3572–3585, Jun. 2021.
- [14] J. Li, Y. Zhou, and H. Chen, "Age of information for multicast transmission with fixed and random deadlines in IoT systems," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8178–8191, Mar. 2020.
- [15] F. Chiariotti, O. Vikhrova, B. Soret, and P. Popovski, "Peak age of information distribution for edge computing with wireless links," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3176–3191, May 2021.
- [16] N. Akar, O. Doğan, 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 Transactions on Communications*, vol. 68, no. 9, pp. 5661–5672, Jun. 2020.
- [17] P. D. Mankar, M. A. Abd-Elmagid, and H. S. Dhillon, "Spatial distribution of the mean peak age of information in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4465–4479, Feb. 2021.
- [18] Z. Li, L. Xiang, and X. Ge, "Age of Information modeling and optimization for fast information dissemination in vehicular social networks," *IEEE Transactions on Vehicular Technology*, Mar. 2022.
- [19] C. Kim and A. K. Agrawala, "Analysis of the fork-join queue," *IEEE Transactions on Computers*, vol. 38, no. 2, pp. 250–255, Feb. 1989.
- [20] W. R. KhudaBukhsh, A. Rizk, A. Frömmgen, and H. Koeppl, "Optimizing stochastic scheduling in fork-join queueing models: Bounds and applications," in *Conference on Computer Communications (INFOCOM)*. IEEE, May 2017, pp. 1–9.
- [21] G. Joshi, E. Soljanin, and G. Wornell, "Efficient redundancy techniques for latency reduction in cloud systems," *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, vol. 2, no. 2, pp. 1–30, Feb. 2017.
- [22] N. B. Shah, K. Lee, and K. Ramchandran, "When do redundant requests reduce latency?" *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 715–722, Dec. 2015.
- [23] Y. Sun, C. E. Koksal, and N. Shroff, "On delay-optimal scheduling in queueing systems with replications," *ArXiv*, vol. abs/1603.07322, Mar. 2016.
- [24] A. Rizk, F. Poloczek, and F. Ciucu, "Stochastic bounds in fork-join queueing systems under full and partial mapping," *Queueing Systems*, vol. 83, no. 3, pp. 261–291, Aug. 2016.
- [25] M. Fidler and Y. Jiang, "Non-asymptotic delay bounds for (k, l) fork-join systems and multi-stage fork-join networks," in *35th Annual International Conference on Computer Communications (INFOCOM)*. IEEE, Apr. 2016.
- [26] M. Bastopcu and S. Ulukus, "Age of Information for updates with distortion: Constant and age-dependent distortion constraints," *IEEE/ACM Transactions on Networking*, Aug. 2021.
- [27] S. Hu and W. Chen, "Balancing data freshness and distortion in real-time status updating with lossy compression," in *Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, Jul. 2020, pp. 13–18.
- [28] B. Buyukates, M. Bastopcu, and S. Ulukus, "Optimal selective encoding for timely updates with empty symbol," in *International Symposium on Information Theory (ISIT)*. IEEE, Jun. 2020, pp. 1794–1799.
- [29] B. Buyukates and S. Ulukus, "Timely distributed computation with stragglers," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5273–5282, Jun. 2020.
- [30] R. Talak and E. H. Modiano, "Age-delay tradeoffs in queueing systems," *IEEE Transactions on Information Theory*, vol. 67, no. 3, pp. 1743–1758, Mar. 2021.

- [31] A. K. Erlang, "Løsning af nogle problemer fra sandsynlighedsregningen af betydning for de automatiske telefoncentraler," *Elektroteknikeren*, vol. 13, pp. 5–13, Jan. 1917.
- [32] D. Pinotsi and M. A. Zazanis, "Synchronized queues with deterministic arrivals," *Operations Research Letters*, vol. 33, no. 6, pp. 560–566, Nov. 2005.
- [33] F. Baccelli and P. Brémaud, *Palm probabilities and stationary queues*, ser. Lecture Notes in Statistics. Springer Verlag, Dec. 2012, vol. 41.
- [34] A. Prudnikov, Y. A. Brychkov, and O. I. Marichev, *Integrals and series, Volume 1: Elementary functions*. Gordon&Breach Scientific Publishing, New York, 1986.
- [35] M. Wiper, "Bayesian analysis of $Er/M/1$ and $Er/M/c$ queues," *Journal of Statistical Planning and Inference*, vol. 69, no. 1, pp. 65–79, Jun. 1998.
- [36] L. Flatto and S. Hahn, "Two parallel queues created by arrivals with two demands I," *SIAM Journal on Applied Mathematics*, vol. 44, no. 5, pp. 1041–1053, Oct. 1984.
- [37] L. Flatto, "Two parallel queues created by arrivals with two demands II," *SIAM Journal on Applied Mathematics*, vol. 45, no. 5, pp. 861–878, Oct. 1985.



Petar Popovski (S'97–A'98–M'04–SM'10–F'16) is a Professor at Aalborg University, where he heads the section on Connectivity and a Visiting Excellence Chair at the University of Bremen. He received his 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 a Fellow of the IEEE. 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), the Danish Telecommunication Prize (2020) and Villum Investigator Grant (2021). He is a Member at Large at the Board of Governors in IEEE Communication Society, Vice-Chair of the IEEE Communication Theory Technical Committee and IEEE Transactions on Green Communications and Networking. He is currently an Editor-in-Chief of IEEE Journal on Selected Areas in Communications. Prof. Popovski was the General Chair for IEEE SmartGridComm 2018 and IEEE Communication Theory Workshop 2019. His research interests are in the area of wireless communication and communication theory. He authored the book "Wireless Connectivity: An Intuitive and Fundamental Guide," published by Wiley in 2020.



Federico Chiariotti (S'15–M'19) is currently an assistant professor at the Department of Electronic Systems, Aalborg University, Denmark. He received his Ph.D. in information engineering in 2019 from the University of Padova, Italy. He received the bachelor's and master's degrees in telecommunication engineering (both *cum laude*) from the University of Padova, in 2013 and 2015, respectively. He has authored over 40 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.



Beatriz Soret (M'11) received her M.Sc. and Ph.D. degrees in Telecommunications from the University of Malaga, Spain, in 2002 and 2010, respectively. She is currently an associate professor at the Department of Electronic Systems, Aalborg University, and a Senior Research Fellow at the Communications Engineering Department, University of Malaga. Her research interests include LEO satellite communications, AoI and semantic communications, and 5G and 6G systems.