Status updates through M/G/1/1 queues with HARQ

Elie Najm
LTHI, EPFL, Lausanne, Switzerland
Email: elie.najm@epfl.ch

Roy Yates
ECE Dept., Rutgers University, USA
Email: ryates@rutgers.edu

Emina Soljanin
ECE Dept., Rutgers University, USA
Email: emina.soljanin@rutgers.edu

***Abstract*—We consider a system where randomly generated updates are to be transmitted to a monitor, but only a single update can be in the transmission service at a time. Therefore, the source has to prioritize between the two possible transmission policies: preempting the current update or discarding the new one. We consider Poisson arrivals and general service time, and refer to this system as the M/G/1/1 queue. We start by studying the average status update age and the optimal update arrival rate for these two schemes under general service time distribution. We then apply these results on two practical scenarios in which updates are sent through an erasure channel using (a) an infinite incremental redundancy (IIR) HARQ system and (b) a fixed redundancy (FR) HARQ system. We show that in both schemes the best strategy would be not to preempt. Moreover, we also prove that, from an age point of view, IIR is better than FR.**

I. Introduction

Previous work on status update ([1]–[6]) used an Age of Information (AoI) metric in order to assess the freshness of randomly generated updates sent by one or multiple sources to a monitor through the network. In these papers, updates are assumed to be generated according to a Poisson process and the main metric used to quantify the *age* is the time average age (which we will call average age) given by

$$\Delta = \lim_{\tau\to\infty} \frac{1}{\tau}\int_0^\tau \Delta(t)\mathrm{d}t, \tag{1}$$

where $\Delta(t)$ is the instantaneous age of the last successfully received update. If this update was generated at time $u(t)$ then its *age* at time t is $\Delta(t) = t - u(t)$. When the system is idle or an update is being transmitted then the instantaneous age increases linearly with time, as depicted in Fig. 1. Once an update generated at time t_i is received by the monitor at t_i', $\Delta(t)$ drops to the value $t_i' - t_i$. This results in the sawtooth sample path seen in Fig. 1.

In this paper, we assume updates are generated according to a Poisson process with rate λ, but the system can handle only one update at a time without any buffer to store incoming updates. This means that whenever a new update is generated and the system is busy, the transmitter has to make a decision: does it give higher priority to the new update or to the one being transmitted? In other words, does it preempt or not? It has been shown that for exponential update service times, preemption ensures the lowest average age [2]. However, the work in [5] suggests that under the assumption of gamma distributed service time, preemption might not be the best policy.

This work answers the previous question when we assume updates are sent through a symbol erasure channel with erasure rate δ, while using hybrid ARQ (HARQ) protocols to combat erasures. Two HARQ protocols, introduced in [7], are studied: (a) infinite incremental redundancy (IIR) and (b) fixed redundancy (FR). In both cases we assume a generated update contains K information symbols. In IIR, encoding is performed at the physical layer where the K information symbols are encoded using a rateless code. Hence, the transmission of an update continues until $k_s = K$ unerased symbols are received. As for the FR, coding is applied at the physical and packet layer. This means that the update is divided into k_p packets with each packet encoded using an (n_s, k_s)-Maximum Distance Separable (MDS) code. So, in this case, the total number of information symbols is $K = k_p k_s$. At the packet level we apply a rateless code and thus the transmission of an update terminates when k_p unerased packets are received. In order to decode a packet, the receiver needs to wait for n_s encoded symbols. Once received, a packet is declared erased if fewer than k_s symbols are successful. It is worth noting that in this setup we send one symbol per channel use and thus the arrival rate λ is the number of updates generated per channel use. The effect of these schemes on the transmission time of data was studied in [7]. It was shown that FR leads to a slower delivery than IIR. While the main aim of [7] is the successful delivery of every update, in this paper we are ready to sacrifice some updates for fresher information.

The impact of transmission error on the age was also investigated in [8]. In this paper, service time is assumed exponential and another age metric is used: the peak age of information. The authors conclude that, in this setup, preemption with update retransmission achieves the lowest age.

To solve the above problem, we first start by deriving in Section III an expression for the average age under general service time distribution when we choose not to preempt. This model is called M/G/1/1 with blocking. In Section IV, we use the results in the previous Section to compute the average age when we consider the IIR and FR protocols. Sections V and VI follow the same logic but in this case we choose to preempt. This model is called M/G/1/1 with preemption. Finally, Section VII compares the performances of both models for a given HARQ protocol as well as the performance of both protocols given a model. We show that no matter the protocol, prioritizing the current update is better than preempting it. Moreover, in the case of FR, we show that no matter the model and for a fixed arrival rate λ, there exists an optimal codeword length n_s.

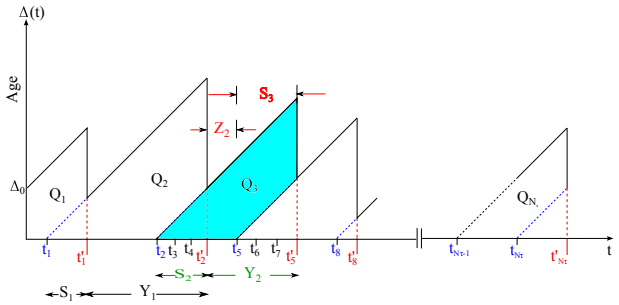

Fig. 1. Variation of the instantaneous age for M/G/1/1 with blocking

II. Preliminaries

It is important to note that in both M/G/1/1 queues, some updates might be dropped. Hence we call the updates that are not dropped, and thus delivered to the receiver, as "successfully received updates" or "successful updates". In addition to that, we also define: (i) I_i to be the true index of the i^{th} successfully received update, (ii) $Y_i = t'_{I_{i+1}} - t'_{I_i}$ to be the interdeparture time between two consecutive successfully received updates, (iii) $X_i = t_{I_i+1} - t_{I_i}$ to be the interarrival time between the successfully transmitted update and the next generated one (which may or may not be successfully transmitted), so $f_X(x) = \lambda e^{-\lambda x}$, (iV) S_{I_i} to be the service time of the I_i^{th} update with distribution $F_S(t)$, (v) T_i to be the system time, or the time spent by the i^{th} successful update in the queue and (vi) $N_\tau = \max\{n; t_{I_n} \leq \tau\}$, the number of successfully received updates in the interval $[0, \tau]$. In our models, we assume the service time S_k of the k^{th} update is independent from the interarrival time random variables $\{X_1, X_2, ..., X_k, ...\}$ and that the sequence $\{S_1, S_2, ...\}$ forms an i.i.d process.

From (1), Fig. 1 and Fig. 3, the average age for both M/G/1/1 queues can be also expressed as the sum of the geometric areas Q_i under the instantaneous age curve. Authors in [2] show that

$$\Delta = \lim_{\tau\to\infty} \frac{N_\tau}{\tau}\frac{1}{N_\tau}\sum_{i=1}^{N_\tau} Q_i = \lambda_e \mathbb{E}(Q_i), \tag{2}$$

where $\lambda_e = \lim_{\tau\to\infty} \frac{N_\tau}{\tau}$ and the second equality is justified by the ergodicity of the system.

III. M/G/1/1 with Blocking

In this setup, a generated update is discarded if it finds the system busy. This means an update is served only if it arrives at an idle system. This concept is depicted in Fig. 1: for instance, the update generated at time t_2 is served since the system is empty at that time. However, the updates generated at times t_3 and t_4 find the system busy and are thus discarded. One important note here is that the system time T_i of the i^{th} successful update is equal to its service time.

A. Average age calculation

Lemma 1. *For an M/G/1/1 blocking system we have,*

$$\lambda_e = \frac{1}{\mathbb{E}(Y)} = \frac{1}{\mathbb{E}(X) + \mathbb{E}(S)}, \tag{3}$$

where Y, X and S are the steady-state counterparts of the variables defined in Section II.

Proof. N_τ is a renewal process with inter-renewal time between two renewals given by the random variable Y. As shown in Fig. 1, the renewal period is the interval:

$$Y_i = Z_i + S_{i+1}. \tag{4}$$

Because each departure leaves the system empty and the interarrival times are memoryless, then the interval Z_i, which is the residual interarrival time until a new update is generated, is independent of Y_{i-1} and it has an exponential distribution. Hence, all the Y_i's are identically distributed and the Z_i's are stochastically equal to the interarrival time X. This proves why N_τ is a renewal process. The claim follows [9]. □

Now we can compute the average age which is given by the following theorem,

Theorem 1. *The average age of an M/G/1/1 system with blocking is*

$$\Delta = \mathbb{E}(S)\left(\frac{\beta}{2}(C_S + 1) + \frac{1}{\beta}\right), \tag{5}$$

where $C_S = \frac{Var(S)}{\mathbb{E}(S)^2}$ is the squared coefficient of variation and $\beta = \frac{\rho}{\rho+1}$ with $\rho = \frac{\mathbb{E}(S)}{\mathbb{E}(X)} = \lambda\mathbb{E}(S)$.

Proof. From (2) we have,

$$\Delta = \lambda_e \mathbb{E}(Q_i).$$

λ_e is given by Lemma 1, therefore we need to compute the average area of the trapezoid Q_i. To do that, notice first that, using a similar argument as the one used in the proof of Lemma 1, the service time S_i and Y_i are independent. Thus,

$$\begin{aligned}\mathbb{E}(Q_i) &= \mathbb{E}\left(\frac{(S_{i-1} + Y_{i-1})^2}{2} - \frac{S_i^2}{2}\right) \\ &= \frac{1}{2}\mathbb{E}\left(Y_{i-1}^2\right) + \mathbb{E}(S_{i-1})\mathbb{E}(Y_{i-1}).\end{aligned} \tag{6}$$

Since we are interested in the steady-state behavior, we will drop the subscript index on the random variables. Hence,

$$\begin{aligned}\mathbb{E}(Q) &= \frac{1}{2}\mathbb{E}\left(Y^2\right) + \mathbb{E}(S)\mathbb{E}(Y) \\ &= \frac{1}{2}\mathbb{E}\left((X+S)^2\right) + \mathbb{E}(S)\mathbb{E}(S+X) \\ &= \frac{1}{2}\left(\mathbb{E}\left(X^2\right) + \mathbb{E}\left(S\right)^2\right) + \frac{1}{2}\mathrm{Var}(S) + 2\mathbb{E}(S)\mathbb{E}(X) \\ &\quad + \mathbb{E}(S)^2 \\ &= \frac{1}{2}\left(\mathbb{E}\left(S\right)^2 + \mathrm{Var}(S)\right) + \mathbb{E}\left(X\right)^2 + 2\mathbb{E}(S)\mathbb{E}(X) \\ &\quad + \mathbb{E}(S)^2 \\ &= \left(\mathbb{E}(X) + \mathbb{E}(S)\right)^2 + \frac{1}{2}\left(\mathbb{E}(S)^2 + \mathrm{Var}(S)\right),\end{aligned} \tag{7}$$

where the third equality is obtained by adding and subtracting $\frac{1}{2}\mathbb{E}(S)^2$ to the second equality, and the fourth equality is obtained by noticing that for the exponential random variable X we have $\mathbb{E}\left(X^2\right) = 2\mathbb{E}(X)^2$. Using (3) and (7), we get (5). □

B. Finding the optimal arrival rate

When the arrival rate of the updates is a parameter that we can control, it is interesting to have an idea on its value that minimizes the average age.

Theorem 2. *For the M/G/1/1 blocking system, the minimum average age Δ^* is achieved for:*

- *If $C_S > 1$, then $\lambda^* = \frac{\beta^*}{(1-\beta^*)\mathbb{E}(S)}$ with $\beta^* = \sqrt{\frac{2}{C_S+1}}$ and*
$$\Delta^* = \mathbb{E}(S)\sqrt{2(C_S+1)}$$
- *If $C_S \leq 1$, $\lambda^* \to \infty$ and $\Delta^* = \mathbb{E}(S)\left(\frac{1}{2}(C_S+1)+1\right)$*

Proof. Setting the derivative of (5) with respect to β to zero, we get:
$$\beta^{*2} = \frac{2}{C_S+1}, \tag{8}$$
where β^* is the optimal value of β. Since $0 \leq \beta^* = \frac{\rho^*}{\rho^*+1} < 1$, C_S has to be strictly bigger than 1 for β^* to exist. In this case, $\beta^* = \sqrt{\frac{2}{C_S+1}}$ and solving for λ we get $\lambda^* = \frac{\beta^*}{(1-\beta^*)\mathbb{E}(S)}$. Using β^* in (5) gives the value of the minimum age Δ^*.

If the service time distribution is such that $C_S \leq 1$, then $\frac{\partial\Delta}{\partial\beta} = -\frac{1}{\beta^2} + \frac{C_S+1}{2} < 0$. However, $\frac{\partial\beta}{\partial\lambda} = \frac{\mathbb{E}(S)}{(\lambda\mathbb{E}(S)+1)^2} \geq 0$. Therefore, $\frac{\partial\Delta}{\partial\lambda} = \frac{\partial\Delta}{\partial\beta}\frac{\partial\beta}{\partial\lambda} < 0$. Thus the average age is a strictly decreasing function of the arrival rate and the minimal average age is obtained as $\lambda \to \infty$. □

IV. M/G/1/1 BLOCKING HARQ SYSTEM

Now, we study the effect of different HARQ policies on the average age when considering an M/G/1/1 queue without preemption. We assume that the updates are sent through a symbol erasure channel with erasure rate δ. Moreover, two HARQ protocols are visited: the infinite incremental redundancy (IIR) and the fixed redundancy (FR).

A. Infinite Incremental Redundancy

In this policy, an update consists of k_s information symbols and is encoded using a rateless code. This means that the monitor needs to receive at least k_s symbols in order to decode the update. The transmission of an update finishes whenever k_s symbols are successfully transmitted. All updates arriving when the system is busy are discarded. Therefore, we define the service time S of an update as the number of channel uses needed for the monitor to receive k_s symbols. Hence, S is distributed as a negative binomial with k_s successes and success probability $1-\delta$.

Theorem 3. *The average age of the M/G/1/1 blocking IIR-HARQ system is:*
$$\Delta_{\textit{NIIR}} = \frac{1}{\lambda} + \frac{k_s}{1-\delta} + \frac{\lambda k_s(k_s+\delta)}{2(1-\delta)(\lambda k_s+1-\delta)}. \tag{9}$$
Moreover, the minimum average age is achieved for $\lambda \to \infty$ and its value is given by,
$$\Delta^*_{\textit{NIIR}} = \frac{3k_s+\delta}{2(1-\delta)} \tag{10}$$

Proof. Since we are using IIR policy then the service time S of each update is distributed as a negative binomial $(k_s, 1-\delta)$, $S \in \{k_s, k_s+1, \dots\}$. In this case the mean and variance of S are given by:
$$\mathbb{E}(S) = \frac{k_s}{1-\delta}, \quad \mathrm{Var}(S) = \frac{k_s\delta}{(1-\delta)^2}. \tag{11}$$
Hence, we compute the quantities ρ, β and C_S present in (5):
$$\rho = \frac{\lambda k_s}{1-\delta}, \quad \beta = \frac{\rho}{\rho+1} = \frac{\lambda k_s}{\lambda k_s+1-\delta}, \quad C_S = \frac{\delta}{k_s}. \tag{12}$$
Using the above expression in (5) and performing some simplifications we get (9).

Moreover, since $\delta \leq 1$ and $k_s \geq 1$, $C_S = \frac{\delta}{k_s} \leq 1$. By Theorem 2, the optimum average age is achieved as $\lambda \to \infty$. Taking the limit on (9) gives (10). □

B. Fixed Redundancy

In this policy, we apply two levels of coding: a packet level and a physical level. Each update consists of k_p packets encoded using a rateless code. This means that the monitor needs to receive k_p decodable packets in order to decode the update. Moreover, each packet contains k_s information symbols and is encoded using a (n_s, k_s)-Maximum Distance Separable (MDS) code. Hence, a packet can be decoded if at least k_s symbols are not erased. Since the packets are being transmitted through a symbol erasure channel with erasure probability δ than the probability for the receiver to be unable to decode a packet is:
$$\begin{aligned}\epsilon_p &= \mathbb{P}(\text{less than } k_s \text{ symbols received})\\ &= \sum_{i=0}^{k_s-1}\binom{n_s}{i}\delta^{n_s-i}(1-\delta)^i.\end{aligned} \tag{13}$$

Theorem 4. *The average age of the M/G/1/1 FR-HARQ blocking system is*
$$\Delta_{\textit{NFR}} = \frac{1}{\lambda} + \frac{n_s k_p}{1-\epsilon_p} + \frac{\lambda n_s^2 k_p(k_p+\epsilon_p)}{2(1-\epsilon_p)(\lambda n_s k_p+1-\epsilon_p)}. \tag{14}$$
Moreover, the minimum average age is achieved as $\lambda \to \infty$ and its value is given by,
$$\Delta^*_{\textit{NFR}} = \frac{3n_s k_p+\epsilon_p}{2(1-\epsilon_p)} \tag{15}$$

Proof. The number M of packets needed to be transmitted to decode an update is distributed as a negative binomial $(k_p, 1-\epsilon_p)$ random variable with k_p successes and success rate $(1-\epsilon_p)$, $M \in \{k_p, k_p+1, \dots\}$. Since the transmission of each packet consumes n_s channel uses then the service time S of

each update is $S = n_s M$. Thus, the mean and variance of S are given by:

$$\mathbb{E}(S) = \mathbb{E}(n_s M) = n_s \mathbb{E}(M) = \frac{n_s k_p}{1-\epsilon_p}, \tag{16}$$

$$\mathrm{Var}(S) = \mathrm{Var}(n_s M) = n_s^2 \mathrm{Var}(M) = \frac{n_s^2 k_p \epsilon_p}{(1-\epsilon_p)^2}. \tag{17}$$

Hence, we compute the quantities:

$$\rho = \frac{\lambda k_p}{1-\epsilon_p}, \quad \beta = \frac{\lambda k_p}{\lambda k_p + 1 - \epsilon_p}, \quad C_S = \frac{\epsilon_p}{k_p}. \tag{18}$$

Using the above expressions in (5) and performing some simplifications we get (14).

Moreover, since $\epsilon_p \leq 1$ and $k_p \geq 1$, $C_S = \frac{\epsilon_p}{k_p} \leq 1$. By Theorem 2, the optimum average age is achieved as $\lambda \to \infty$. From (14) this yields (15). □

V. M/G/1/1 WITH PREEMPTION

In the M/G/1/1 with preemption scenario, any packet being served is preempted if a new packet arrives and the new packet is served instead. In fact, while in the M/G/1/1 with blocking the priority is given to the update being served, in this setup the priority goes to the newly generated update. Moreover, the number of packets in the queue can be modeled as a continuous-time two-state semi-Markov chain depicted in Figure 2.

The 0-state corresponds to empty queue and no packet is being served while the 1-state corresponds to the state where the queue is full and is serving one packet. However, given that the interarrival time between packets is exponentially distributed with rate λ then one spends an exponential amount of time X in the 0-state before jumping with probability 1 to the other state. Once in the 1-state, two independent clocks are started: the service time clock of the packet being served and the rate λ memoryless clock of the interarrival time between the current packet and the next one to be generated. If the memoryless clock ticks first, we stay in the 1-state, otherwise we go back to the 0-state. Hence, the jump from the 1-state to the 0-state occurs with probability $p = \mathbb{P}(S < X)$, where S is a generic service time with distribution $f_S(t)$ and X is a generic rate λ memoryless interarrival time which is independent of S.

The quantity p will play an important role in our derivation, so we will take a closer look at it:

$$p = \int_0^\infty f_S(t)\mathbb{P}(X > t)\ \mathrm{d}t = \int_0^\infty f_S(t) e^{-\lambda t}\ \mathrm{d}t = P_\lambda, \tag{19}$$

where P_λ is the Laplace transform of the service time distribution.

Using Fig. 3 it was shown in [5] that the average age Δ is:

$$\Delta = \lambda_e \mathbb{E}(Q) = \lambda_e \left(\frac{1}{2}\mathbb{E}\left(Y^2\right) + \mathbb{E}(T)\mathbb{E}(Y) \right), \tag{20}$$

where $\lambda_e = \lambda P_\lambda$ is the effective arrival rate, T and Y as defined in Section II. We start with $\mathbb{E}(T)$.

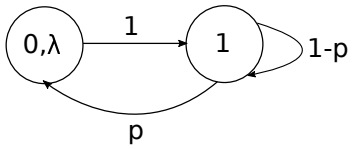

Fig. 2. Semi-Markov chain representing the queue for LCFS with preemption

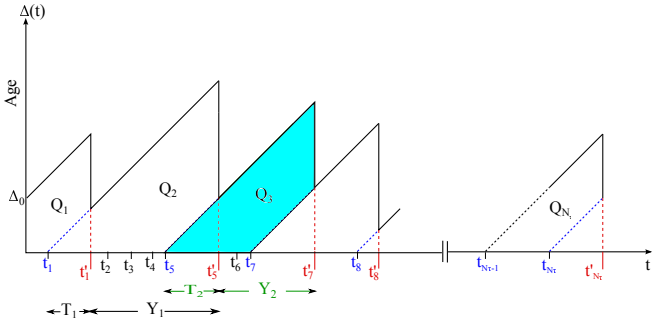

Fig. 3. Variation of the instantaneous age for LCFS with preemption

Lemma 2. *The PDF of the system time T is*

$$f_T(t) = \frac{f_S(t)}{P_\lambda} e^{-\lambda t}. \tag{21}$$

Its expected value is

$$\mathbb{E}(T) = -\frac{1}{P_\lambda}\frac{\partial P_\lambda}{\partial \lambda}. \tag{22}$$

Proof.

$$\begin{aligned} f_T(t) &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(S \in [t, t+\epsilon] | S < X)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(S \in [t, t+\epsilon])}{\epsilon P_\lambda} \mathbb{P}(S < X | S \in (t, t+\epsilon)) \\ &= \frac{f_S(t)}{P_\lambda}\mathbb{P}(X > t) = \frac{f_S(t)}{P_\lambda} e^{-\lambda t}. \end{aligned} \tag{23}$$

Using (21) we calculate the expected value of T:

$$\mathbb{E}(T) = \frac{1}{P_\lambda}\int_0^\infty t f_S(t) e^{-\lambda t} \mathrm{d}t = -\frac{1}{P_\lambda}\frac{\partial P_\lambda}{\partial \lambda}. \tag{24}$$

□

Now we only need to calculate the first and second moments of Y. For that we will derive its moment generating function.

Lemma 3. *The moment generating function of the interdeparture time Y is given by*

$$\phi_Y(s) = \frac{\lambda P_{\lambda - s}}{\lambda P_{\lambda - s} - s}, \tag{25}$$

where $P_{\lambda-s} = \int_0^\infty f_S(t) e^{-(\lambda - s)}\ \mathrm{d}t$.

Proof. From Fig. 3 we can deduce that Y is the shortest time to go from the 0-state back to the 0-state. This means that

$$Y = X + W, \tag{26}$$

where X is exponentially distributed with rate λ and W is

$$W = \begin{cases} T & \text{with probability } p \\ X_1' + T & \text{with probability } (1-p)p \\ X_1' + X_2' + T & \text{with probability } (1-p)^2 p \\ \vdots \end{cases}$$
$$= \sum_{j=0}^{M} X_j' + T, \tag{27}$$

where $X_0' = 0$ and for $j > 0$, X_j' is such that $\mathbb{P}(X_j' < \alpha) = \mathbb{P}(X < \alpha | X < S)$. M, which gives the number of discarded packets before the first successful reception, is a geometric(p) random variable independent of X_j' and T. We start first by deriving the moment generating function of X'.

$$\begin{aligned} f_{X'}(t) &= \lim_{\epsilon \to 0} \frac{\mathbb{P}(X \in [t, t+\epsilon] | S > X)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\mathbb{P}X \in [t, t+\epsilon])}{\epsilon(1-P_\lambda)} \mathbb{P}(S > X | X \in (t, t+\epsilon)) \\ &= \frac{f_X(t)}{1-P_\lambda} \mathbb{P}(S > t) \\ f_{X'}(t) &= [1 - F_S(t)] \frac{\lambda e^{-\lambda t}}{1-P_\lambda}, \end{aligned} \tag{28}$$

where $F_S(t)$ is the cdf of the service time S. Hence,

$$\begin{aligned} \phi_{X'}(s) = \mathbb{E}\left(e^{sX'}\right) &= \int_0^\infty e^{st} \left(1 - F_S(t)\right) \frac{\lambda e^{-\lambda t}}{1-P_\lambda} \, \mathrm{d}t \\ &\overset{(a)}{=} \frac{\lambda}{\lambda - s} \frac{1}{1-P_\lambda} - \frac{\lambda}{1-P_\lambda} \frac{P_{\lambda - s}}{\lambda - s} \\ &= \frac{\lambda(1 - P_{\lambda - s})}{(\lambda - s)(1 - P_\lambda)}, \end{aligned} \tag{29}$$

where (a) is obtained by using integration by parts with $u = 1 - F_S(t)$ and $\frac{\mathrm{d}v}{\mathrm{d}t} = e^{-t(\lambda - s)}$. On the other hand, (21) implies

$$\phi_T(s) = \mathbb{E}\left(e^{sT}\right) = \int_0^\infty \frac{f_S(t)}{P_\lambda} e^{-\lambda t} e^{st} \, \mathrm{d}t = \frac{P_{\lambda - s}}{P_\lambda}. \tag{30}$$

Using (29) and (30), we deduce the moment generating of W,

$$\begin{aligned} \phi_W(s) &= \mathbb{E}\left(e^{s\left(\sum_{i=0}^{M} X_i' + T\right)}\right) \\ &= \mathbb{E}\left(e^{sT}\right) \mathbb{E}\left(\mathbb{E}\left(e^{sX'}\right)^M\right) \\ &= \frac{P_{\lambda - s}}{P_\lambda} \sum_{i=0}^{\infty} \left(\frac{\lambda(1 - P_{\lambda - s})}{(\lambda - s)(1 - P_\lambda)}\right)^i (1 - P_\lambda)^i P_\lambda \\ &= \frac{(\lambda - s) P_{\lambda - s}}{\lambda P_{\lambda - s} - s}. \end{aligned} \tag{31}$$

Using (31) and that $\phi_X = \mathbb{E}\left(e^{sX}\right) = \frac{\lambda}{\lambda - s}$, we get (25) from $\phi_Y(s) = \mathbb{E}\left(e^{sX}\right) \mathbb{E}\left(e^{sW}\right)$. □

Theorem 5. *The average age of an M/G/1/1 system with preemption is given by,*

$$\Delta = \lambda_e \mathbb{E}(Q) = \frac{1}{\lambda P_\lambda}. \tag{32}$$

Proof. Deriving (25) once and twice and setting $s = 0$ gives:

$$\mathbb{E}(Y) = \frac{1}{\lambda P_\lambda} \quad \text{and} \quad \mathbb{E}(Y^2) = \frac{2}{\lambda^2 P_\lambda^2} \left(1 + \lambda \frac{\partial P_\lambda}{\partial \lambda}\right) \tag{33}$$

Using (22) and (33) we get $\mathbb{E}(Q) = \frac{1}{\lambda^2 P_\lambda^2}$. This last expression and the fact that $\lambda_e = \lambda P_\lambda$ give (32). □

In conclusion, for the M/G/1/1 with preemption, the average age depends on the Laplace transform of the service time distribution.

VI. M/G/1/1 WITH PREEMPTION AND HARQ

In this Section we study the effect of different HARQ policies on the average age when considering an M/G/1/1 queue with preemption. Indeed, we assume that the updates are sent through a symbol erasure channel with erasure rate δ. Moreover, two HARQ models are visited: the infinite incremental redundancy (IIR) and the fixed redundancy (FR).

A. Infinite Incremental Redundancy

In this setup, the transmission of an update finishes whenever one of these events happen first: (i) k_s symbols are successfully transmitted, or (ii) a new update is generated. Hence the following theorem.

Theorem 6. *The average age of an M/G/1/1 with preemption system when using the IIR policy is given by,*

$$\Delta_{\mathit{PIIR}} = \frac{1}{\lambda} \left(\frac{e^\lambda - \delta}{1 - \delta}\right)^{k_s}. \tag{34}$$

Moreover, Δ_{PIIR} has a minimum and the arrival rate λ^* that achieves it should satisfy the condition

$$\lambda^* \leq \frac{1}{k_s}. \tag{35}$$

The minimum age Δ^*_{PIIR} can be lower bounded using

$$\Delta^*_{\mathit{PIIR}} \geq \frac{1}{\lambda_{\mathit{IIR}}} \left(1 + \frac{\lambda_{\mathit{IIR}}}{1 - \delta}\right)^{k_s}, \tag{36}$$

where $\lambda^* \approx \lambda_{\mathit{IIR}} = \frac{1 - k_s + \sqrt{(k_s+1)^2 - 4k_s\delta}}{2k_s}$.

Proof. Under the IIR policy, the service time S of each update is distributed as a negative binomial $(k_s, 1-\delta)$, $S \in \{k_s, k_s + 1, \dots\}$. In this case the moment generating function of S is given by:

$$\phi_S(s) = \mathbb{E}\left(e^{sS}\right) = \left(\frac{1 - e^s \delta}{e^s(1 - \delta)}\right)^{-k_s}. \tag{37}$$

Noting that $P_\lambda = \phi_S(-\lambda)$ and using (32) and (37), we get (34). To prove condition (35) we differentiate Δ_{PIIR} with respect to λ and equate it to zero. This yields

$$-\frac{1}{\lambda}\left(\frac{e^\lambda - \delta}{1 - \delta}\right) + \frac{k_s e^\lambda}{1 - \delta} = 0. \tag{38}$$

Thus, to satisfy (38) we need

$$e^\lambda(k_s\lambda - 1) = -\delta. \tag{39}$$

Since $0 \leq \delta \leq 1$, (39) implies that $k_s\lambda - 1 \leq 0$. Hence (35) holds. Moreover, since $\lambda > 0$, we have that $e^{\lambda} > 1 + \lambda$. This means that if λ^* minimizes Δ_{PIIR}, then

$$\Delta^*_{\text{PIIR}} = \Delta_{\text{PIIR}}(\lambda^*) > \frac{1}{\lambda^*}\left(1 + \frac{\lambda^*}{1-\delta}\right)^{k_s}. \quad (40)$$

Finally, in order to obtain λ^* one needs to solve equation (39) which does not have a simple closed form expression. As an alternative, we can make the small λ approximation $e^{\lambda^*} \approx 1 + \lambda^*$. In this case, (39) reduces to

$$(1+\lambda)(k_s\lambda - 1) = -\delta. \quad (41)$$

This is a quadratic equation whose only positive root is given by

$$\lambda_{IIR} = \frac{1 - k_s + \sqrt{(k_s+1)^2 - 4k_s\delta}}{2k_s}.$$

To obtain (36), we replace λ^* by λ_{IIR} in (40). □

Since $\lambda^* \leq \frac{1}{k_s} \leq 1$, the lower bound in (36) becomes a tight approximation of the average age for typical values of k_s.

B. Fixed Redundancy

In this case also the transmission of an update is terminated whenever one of these events happen first: (i) k_p packets are successfully transmitted, or (ii) a new update is generated. As in the M/G/1/1 blocking system, we define the packet erasure probability $\epsilon_p = \sum_{i=0}^{k_s-1} \binom{n_s}{i} \delta^{n_s-i}(1-\delta)^i$.

Theorem 7. *The average age of the information for an M/G/1/1 with preemption system using the FR policy is given by,*

$$\Delta_{PFR} = \frac{1}{\lambda}\left(\frac{1 - e^{-\lambda n_s}\epsilon_p}{e^{-\lambda n_s}(1-\epsilon_p)}\right)^{k_p}. \quad (42)$$

Moreover, Δ_{PFR} has a minimum and the arrival rate λ^* that achieves it should satisfy the condition

$$\lambda^* \leq \frac{1}{n_s k_p}. \quad (43)$$

The minimum age Δ^*_{PIIR} can be lower bounded using

$$\Delta^*_{PFR} \geq \frac{1}{\lambda_{FR}}\left(1 + \frac{\lambda_{FR} n_s}{1-\epsilon_p}\right)^{k_p}, \quad (44)$$

where $\lambda^* \approx \lambda_{FR} = \frac{1-k_p+\sqrt{(k_p+1)^2-4k_p\epsilon_p}}{2n_s k_p}$.

Proof. The number M of packets needed to be transmitted to decode an update is distributed as a negative binomial $(k_p, 1-\epsilon_p)$ random variable with k_p successes and success rate $(1-\epsilon_p)$, $M \in \{k_p, k_p+1, \dots\}$. Since the transmission of each packet consumes n_s channel uses, the service time S of each update is $S = n_s M$. Thus, the moment generating function of S is:

$$\phi_S(s) = \mathbb{E}\left(e^{sn_sM}\right) = \phi_M(n_s s) = \left(\frac{e^{n_s s}(1-\epsilon_p)}{1 - e^{n_s s}\epsilon_p}\right)^{k_p}. \quad (45)$$

Using (32), the fact that $P_\lambda = \phi_S(-\lambda)$ and the above expression we obtain (42).

To prove condition (43) we differentiate Δ_{PFR} with respect to λ and equate it to zero, yielding

$$-\frac{1}{\lambda}\left(\frac{e^{\lambda n_s} - \epsilon_p}{1-\epsilon_p}\right) + \frac{k_p n_s e^{\lambda n_s}}{1-\epsilon_p} = 0. \quad (46)$$

Thus, to satisfy (46) we need

$$e^{\lambda n_s}(k_p n_s \lambda - 1) = -\epsilon_p. \quad (47)$$

Since $0 \leq \epsilon_p \leq 1$, (47) implies that $k_p n_s \lambda - 1 \leq 0$. Hence (43) holds.

As in the proof for Theorem 6, here also we have:

$$\Delta^*_{\text{PFR}} = \Delta_{\text{PFR}}(\lambda^*) > \frac{1}{\lambda^*}\left(1 + \frac{n_s\lambda^*}{1-\epsilon_p}\right)^{k_p}. \quad (48)$$

Finally, also as in the proof for Theorem 6, we approximate the real value of λ^* by solving the quadratic equation

$$(1 + \lambda n_s)(k_p n_s \lambda - 1) = -\epsilon_p. \quad (49)$$

The only positive root is given by

$$\lambda_{\text{FR}} = \frac{1 - k_p + \sqrt{(k_p+1)^2 - 4k_p\epsilon_p}}{2n_s k_p}.$$

To obtain (44), we replace λ^* by λ_{FR} in (48). □

Since $n_s\lambda^* \leq \frac{1}{k_p} \leq 1$, the lower bound in (44) becomes a tight approximation for typical values of k_p.

VII. Numerical results

In this section, we first compare the two HARQ policies, IIR and FR, for the M/G/1/1 with and without preemption. Then, for each HARQ policy, we compare the performances of the two M/G/1/1 schemes. Moreover, for the simulation results discussed in this section, we assume the following setting: a symbol erasure channel with erasure rate $\delta = 0.2$ and each update in IIR-HARQ and FR-HARQ contain $K = 100$ information symbols. So for IIR-HARQ we have $f_s = 100$ while for FR-HARQ, we assume each update is divided into $k_p = K/k_s$ packets where each packet is encoded using an MDS-(k_s, n_s) code.

We first start analyzing the M/G/1/1 system with preemption. Fig. 4 shows the average age for different values of k_s around its minimum point. As we can notice, if we choose the optimum n_s for a fixed k_s and range of λ then the average age decreases as the number of packets per update decreases. In fact, the black curve which corresponds to $k_p = 1$ has the lowest average age around its minimum, followed by the blue curve associated with $k_p = 5$ and the worst performance is for the system with $k_p = 10$. Fig. 4 also confirms the results in Theorem 6 and Theorem 7 saying that Δ_{PIIR} and Δ_{PFR} achieve a minimum at a small value of λ. This figure also suggests that no matter how we choose k_s and n_s, IIR outperforms FR. The values of n_s chosen in Fig. 4 are such that they minimize the average age for a given δ and k_s. The existence of such optimum packet length in FR can be deduced from Fig. 5.

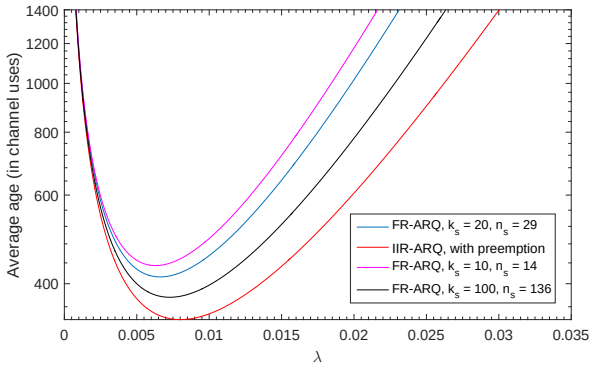

Fig. 4. Comparing the performance of the FR-HARQ for the M/G/1/1 with preemption scheme when varying the number of information symbols in each packet. We assume the update has 100 information symbols, $\delta = 0.2$, $k_p = 100/k_s$. n_s is chosen to minimize the average age.

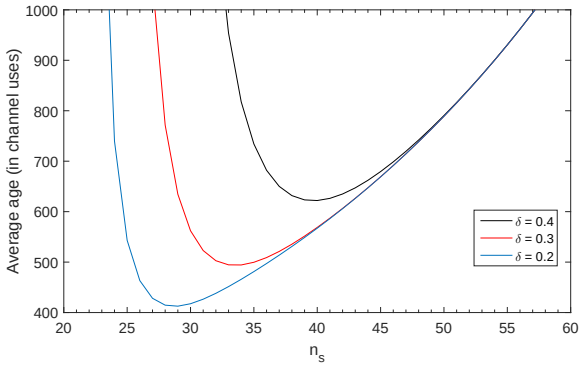

Fig. 5. Average age with respect to codeword length for the M/G/1/1 with preemption scheme with FR-HARQ. We assume the update has 100 information symbols, $\lambda = 0.0066$, $k_s = 20$ and $k_p = 100/k_s$.

Here we set $\lambda = 0.0066$, which minimizes the average age for $\delta = 0.2$, and $k_s = 20$. Fig. 5 can be explained using the lower bound (44): for a given λ, as n_s gets large, $\epsilon_p \to 0$ and the lower bound will be increasing with n_s since $\left(1 + \frac{n_s \lambda^*}{1-\epsilon_p}\right) > 1$. However, for n_s close to k_s, $\epsilon_p \to 1$ which also increases this lower bound. Thus, the packet length should be neither too small (equal to k_s) nor too large. As it is expected, Fig. 5 also shows that the optimal packet length n_s increases as the erasure rate δ increases.

The above results concerning the M/G/1/1 system with preemption apply also for the M/G/1/1 blocking system as it can be seen in Fig. 6 and 7. However, some differences need to be noted. (i) Fig. 6 confirms the results of Theorems 3 and 4 that the average age is a decreasing function of λ. (ii) Fig. 6 shows that for any value of λ, increasing the number of packets per update increases the average age. (iii) Fig. 7 shows the existence of an optimal packet length n_s for a given δ, λ and k_s.

Finally, we compare the performance of the M/G/1/1 with preemption and the M/G/1/1 blocking systems for each one of the HARQ policies. In both cases, Fig. 8 shows that the M/G/1/1 blocking system performs better than its counterpart for all values of λ.

VIII. Conclusion

In this paper we studied the M/G/1/1 system along with the possible update management policies it presents: preempting the current update or discarding the newly generated one. We derived general expressions for their average age and used this result to compute the average age when considering a practical scenario: updates are sent over a symbol erasure channel using two different HARQ protocols, IIR and FR. In both cases, prioritizing the current update being sent and not preempting it turned out to be the best strategy. Moreover, as it is expected, the IIR protocol gives better performance from an age point of view than FR. Finally, we argued through simulations that for the FR protocol, ensuring reliable delivery of every update packet (by using large codeword length n_s) doesn't achieve the optimal average age.

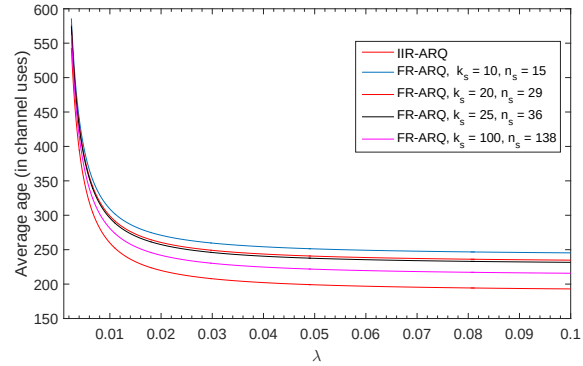

Fig. 6. Comparing the performance of the FR-HARQ for the M/G/1/1 without preemption scheme when varying the number of information symbols in each packet. We assume the update has 100 information symbols, $\delta = 0.2$, $k_p = 100/k_s$. n_s is chosen to minimize the average age.

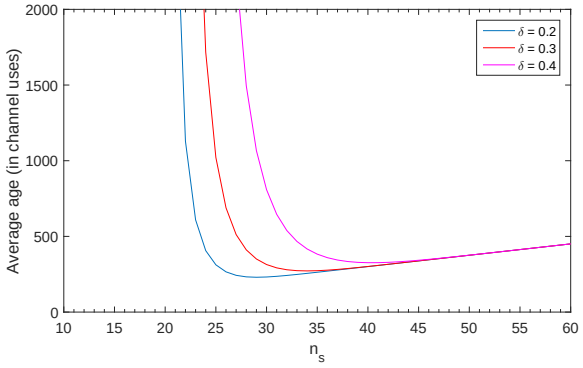

Fig. 7. Average age with respect to codeword length for the M/G/1/1 without preemption scheme with FR-HARQ. We assume the update has 100 information symbols, $\lambda = 1$, $k_s = 20$ and $k_p = 100/k_s$.

References

[1] S. Kaul, R. D. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. INFOCOM*, 2012.

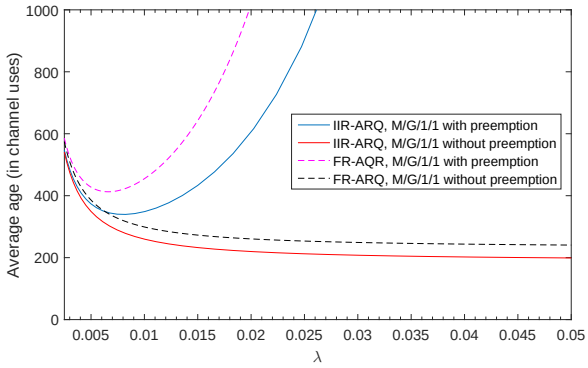

Fig. 8. Comparing the performance of the two M/G/1/1 schemes when using IIR and FR. We assume the update has 100 information symbols and $\delta = 0.2$.

[2] ——, "Status updates through queues," in *Conf. on Information Sciences and Systems (CISS)*, Mar. 2012.

[3] M. Costa, M. Codreanu, and A. Ephremides, "On the age of information in status update systems with packet management," *IEEE Trans. Info Theory*, vol. 62, no. 4, pp. 1897–1910, April 2016.

[4] C. Kam, S. Kompella, and A. Ephremides, "Age of information under random updates," in *Proc. IEEE Int'l. Symp. Info. Theory*, 2013.

[5] E. Najm and R. Nasser, "The age of information: The gamma awakening," in *Proc. IEEE Int'l. Symp. Info. Theory*, 2016, pp. 2574–2578.

[6] R. D. Yates and S. Kaul, "Real-time status updating: Multiple sources," in *Proc. IEEE Int'l. Symp. Info. Theory*, Jul. 2012.

[7] M. Heindlmaier and E. Soljanin, "Isn't hybrid ARQ sufficient?" in *Communication, Control, and Computing (Allerton), 52nd Annual Allerton Conference on*. IEEE, 2014, pp. 563–568.

[8] K. Chen and L. Huang, "Age-of-information in the presence of error," in *Proc. IEEE Int'l. Symp. Info. Theory*, 2016, pp. 2579–2584.

[9] S. M. Ross, *Stochastic Processes (Wiley Series in Probability and Statistics)*, 2nd ed. Wiley, Feb. 1995.