Performance Analysis of Queueing Systems with Systematic Packet-Level Coding

†Giuseppe Cocco, †Tomaso de Cola, *Matteo Berioli
†German Aerospace Center – DLR, Oberpfaffenhofen, 82234 Wessling, Germany
*TriaGnoSys GmbH, 82234 Wessling, Germany
{giuseppe.cocco, tomaso.decola}@dlr.de
matteo.berioli@triagnosys.com

***Abstract*—We study a queueing system operated with packet level coding. More specifically, we derive a closed-form approximation for the queueing delay as well as an expression for the decoding delay of a system operated with systematic network coding. Unlike previous works, the delay is considered on a per-packet basis rather than per-block, thus taking into account the low-latency property of systematic codes. Furthermore we study the trade-off between the coding gain and the decoding delay for finite block lengths.**

I. Introduction

Today's communication networks are demanded to provide reliable, high data-rate and ubiquitous services. In particular, matching requirements of Quality of Service (QoS) such as delay and packet loss rate (PLR) plays a main role in satisfying the Quality of Experience (QoE) figures of merit demanded by each user for the different services being offered.

The massive deployment of mobile networks brings about an intrinsic channel unreliability due to fading, which often can not be compensated at the physical layer due to lack of up-to-date information about the channel state at the transmitter. This is common in networks with large propagation delays such as mobile satellite networks. A possible approach to overcome packet losses over unreliable links is the application of packet erasure codes. Systematic codes are particularly appealing since they can significantly decrease the decoding delay with respect to non-systematic codes. This is thanks to the fact that a systematic packet that is received correctly is readily available to the upper layers and, thus, does not need to wait for the whole code block to be received, unlike non-systematic codes[1]. Thanks to such characteristic, systematic packet level codes have been considered in several standards such as 3GPP, DVB-H and DVB-SH [1].

The drawback of packet level codes is the increase in delay, which is particularly significant in case of high probability of packet loss and when long code blocks are used. There are well established results in the performance of codes for erasure channels. Among such results is the possibility to asymptotically achieve the channel capacity (given by the complement to 1 of the packet erasure rate) as the code block length goes to infinity. However, the interaction of delay and packet loss rate for queueing systems in which packet level codes of finite length are used is still subject of intensive studies.

[1]This is the case if in-order reception is not required, which is assumed in this paper

In particular, network coding (NC) has been extensively investigated over the last decade as a complementary tool to improve the robustness of data communication over error-prone links [2] and to improve the reaction of those higher layer protocols (e.g., TCP) that are more sensitive to packet erasures [3]. Additionally, network coding has been also investigated in hybrid Automatic Repeat reQuest (ARQ) strategies [4], so that additional redundancy is generated in case the decoding procedure at the receiver is not successful. The performance benefits of ARQ could be penalised by large latencies which are typical, for instance, of geostationary satellite systems.

Several studies address the problem of delay and packet loss resulting from the use of network coding as done in [5]. Reference [6] addresses the problem in terms of network flow optimisation for multicast sessions; a similar problem is addressed in [7] where the QoS requirements are the inputs for a routing problem optimisation. The case of network coding applied to service classes is considered in [8] where a queue model is introduced to analyse the delay performance. The study of the buffer occupation during coding operations is explored in [9], which develops a theoretical framework to discuss the performance implications in terms of packet loss and delay. In [10] a queueing model for random linear network coding (RLNC) is proposed under the assumption of Bernoulli arrivals and considering feedback from the receiver. A similar problem is addressed in [11]. In [12] and [13] the delay analysis for a multicast system using RLNC is presented, while a duplex communication model is studied in [14] and [15]. In [16] a delay analysis within a RLNC approach for streaming application and feedback from the receivers is carried out. In [17] systematic network coding is applied in time division duplexing channels, showing a reduced complexity with respect to traditional RLNC while achieving the same asymptotic performances in terms of PLR. In all these works the block decoding delay is used as performance metric. In [17], in which a systematic code is considered, the delay is derived on a per-block basis, and the systematicity of the code is only introduced to decrease complexity rather than decrease the delay. Moreover, in most of these works the transmitter keeps sending redundancy until all packets are decoded, leveraging on feedback from the receiver and leading to a reliable (i.e., error free) system.

Unlike most of these works, we consider systematic network coding, in which each packet is transmitted before the encoding takes place. The main difference of our work with

respect to the previous ones, and particularly [17], is that a per-packet delay rather than block decoding delay is used as performance metric. This particularly suits the study of systematic codes, since the main advantage of such code is the fact that packets correctly received do not experience further delays apart from queueing, transmission and propagation. No feedback is assumed from the receiver. Another novel element in our model is the fact that redundancy packets are generated and transmitted back-to-back once all systematic packets within a block have been transmitted. The transmission of redundancy in queueing systems is usually modeled in literature adopting the model of server with vacation, which has been extensively studied. The novel part in our model lies in the fact that the vacation period starts deterministically after all systematic packets in a block have been transmitted, i.e., it is periodic in the number of systematic packets transmitted. Note that such periodicity can not be observed in the time domain due to the random inter-arrival times of the packets, which makes the analysis non trivial. This makes the system different from an M/G/1 and from an M/G/1 with vacation. We derive an approximated expression on the queueing delay, which gives interesting insights on the behavior of the system and allows for a deep understanding of the impact of important parameters such as code block length, code rate and network load, on the delay. Furthermore, we study the joint behaviour of delay and PLR for generic code block sizes and rates.

The remainder of this paper is structured as follows. In Section II the system model is presented. Section III illustrates the proposed theoretical model, describing the delay and packet loss analysis. A numerical validation and evaluation of the derived analytical formulas is presented in Section IV. Finally, Section V summarises the main outcome of this study and gives an outlook of the possible extensions of this work.

II. System Model

Let us consider a queueing system with Poisson arrivals at rate λ packets per second. Packet lengths and link transmission speed are assumed to be fixed, which implies a deterministic transmission (service) time. We denote the service time as $T_{tx} = \frac{1}{\mu}$, where T_{tx} is the transmission time for a packet on the physical link and μ is the service rate measured in packets per seconds. At the beginning of transmissions the transmitter sends out the first K packets upon arrival and then generates and transmits $N-K$ redundancy packets. Once the redundancy transmission has finished, the next K systematic packets are transmitted and so on. Redundancy packets are created applying RLNC over an extended Galois field of size 2^n on the K systematic packets. n is assumed to be large enough such that the receiver can decode with high probability the whole block of K source packets upon reception of any K packets among the N transmitted. We call a set of K systematic packets for which redundancy is generated an NC block. Although the system considered resembles an M/G/1 (or M/D/1, since the service time is deterministic) queue with vacations, there is a subtle difference that impedes the applications of the well known formulas for the delay of such systems [18]. The difference lies in the fact that in the considered system the server goes on vacation mode and stays in such mode for a deterministic period corresponding to $N-K$ transmission slots of duration T_{tx} after exactly K packets are transmitted, while in a traditional system with vacation the vacation period starts at the end of each busy period and has a duration which is a random variable drawn according to some probability distribution. Although many variants of the model of server with vacation have been studied in the past [19] [20], up to our knowledge the system considered here has not yet been characterized in depth. The model for the transmitter side is depicted in Fig. 1. The channel between the transmitter and the receiver is modeled as en erasure channel with erasure probability $\varepsilon = 1-p$. If a systematic packet is correctly received (i.e., there is no erasure) it does not incur in any decoding delay. On the other hand, if a packet is lost, the receiver needs to receive enough redundancy in order to decode the whole block to which the erased packet belongs. In this case the decoding delay is equal to the block decoding delay. Since no feedback is allowed in the system and finite length code blocks are considered, there is a non zero probability that a packet can not be decoded. In this case the decoding delay is equal to the time needed for the transmission of a whole block of N packets. The block diagram for channel and receiver is shown in Fig. 2.

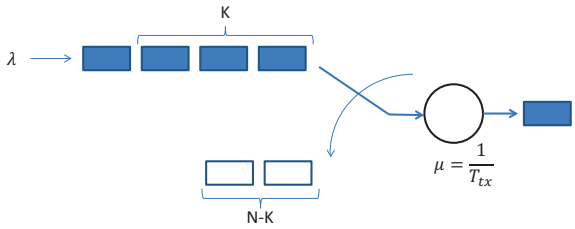

Fig. 1. Transmitter side. Packets arrive according to a Poisson arrival process with rate λ packets per second. A packet is transmitted as soon as it reaches the head of the queue. Transmission time is fixed and equal to T_{tx}. Every $K=3$ systematic packets transmitted, the server (transmitter) switches to a second queue where $N-K=2$ redundancy packets are generated and transmitted back to back. The rate of the system in the picture is $R=K/N=2/3$.

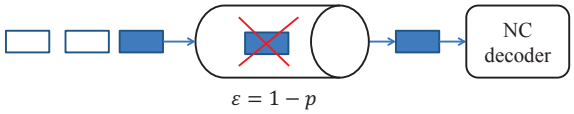

Fig. 2. Channel and receiver side. If a systematic packet is received (on the right) its decoding delay is equal to zero. If an erasure occurs on a systematic packet (center of the picture) the decoding delay for that packet is equal to the block delay. Since no feedback is allowed in the system there is a nonzero probability that an erased packet can not be decoded. In this case the decoding delay for the erased packet is that needed for the transmission of an NC block.

III. Delay Analysis

We define the total delay as the time between the arrival of a packet in the transmitter queue and the instant in which it is either received correctly at the receiver or it is not received correctly but all the redundancy has been transmitted. We denote the average total delay as

$$D_{\text{tot}} = W + G, \tag{1}$$

where W is the average delay introduced by the transmitter side, while G is the delay introduced by the receiver side. W accounts for queueing delay and transmission delay, while G accounts for propagation delay and decoding delay.

A. Transmitter Side

In order to derive the delay introduced by the transmitter side we follow the approach used in [21, Section 3].

Let us consider the instant in which packet i enters the queue at time t finding $N_i(t)$ packets standing between itself and the server, plus an additional packet that is currently being served. The waiting time in the queue for packet i (W_i) is given by [21]

$$W_i \quad = \quad S_i + \sum_{j=i-N_i(t)}^{i-1} X_j + I_i, \tag{2}$$

where S_i is the residual service time of the packet being served, $X_j = T_{\text{tx}}$ is the service time for packet j and I_i is the time needed to send the amount of redundancy relative to the packets already in the queue (or eventually the residual redundancy for the previous block) that will be transmitted before i. I_i depends on whether, at arrival time t, the packet at the server is a systematic or a redundancy packet. It is difficult to derive an exact expression for the waiting time in the queue due to the fact that the number of packets in the queue at a given instant and the position of the packet being served (either systematic or redundancy, if any) within its NC block are not independent. To see this, consider the case in which a packet arrives while the last packet of a redundancy block is being served. During the past $(N-K-1)T_{\text{tx}}$ seconds no systematic packet has being served. This increases the likelihood to find more packets in the queue with respect to the case in which a packet arrives while the K-th systematic packet of a block is being served, since in this case the last $K-1$ packets processed by the server were systematic. Taking this into account, in the following we derive an approximated expression for the delay.

The amount of redundancy an arrival i would find is made up of two parts. One depends only on the number of packets already in the queue and is equal to $(N-K)\left\lfloor \frac{N_i(t)}{K} \right\rfloor$. The second part depends on whether the packet being served (if any) at the time of arrival is systematic or is a redundancy one as well as on the position of the packet in the NC block (i.e., the position in the K-packets block to which a packet belongs). Let us derive an approximation for the second part in case a systematic packet is found at arrival. To see that the second term depends on the position of the packet under service, let us consider a simple example in which a new arrival i finds, at time t, a systematic packet at the server and $N_i(t) = 1$ packets standing between itself and the server. Let us also assume that $K = 3$, and thus $(N-K)\left\lfloor \frac{N_i(t)}{K} \right\rfloor = 0$. In this case, if i is the last packet of its NC block, it will not experience any queueing delay due to redundancy, i.e., it will be transmitted right after the packet standing in front of him. On the other hand, if the newly arrived packet occupies the first position in an NC block, it will have to wait for the transmission of the packet in front of him, which is the last one of the previous NC block, plus the redundancy of the previous block. Averaging across the packet positions within an NC block and assuming that these are independent of the number of packets in the queue[2], the second term of the delay due to redundancy in case a systematic packet is being served at arrival time t can be approximated as:

$$I_i^{\text{sys}}(t) \quad \simeq \quad T_{\text{tx}}(N-K)\frac{\text{mod}\ (N_i(t),K)+1}{K} \tag{3}$$

$\text{mod}(x,y)$ being the rest of the division between x and y. The approximation stems from the fact that, as previously mentioned, $N_i(t)$ and the position of the packet under service within its own block are not independent. The term $\frac{\text{mod}\ (N_i(t),K)+1}{K}$ in the equation takes into account the particular phase in the NC block of the arriving packet. The derivation is straightforward once one notices that, on average, an arriving packet has the same probability $(1/K)$ of being in any of the K possible positions within its NC block. By the property Poisson Arrivals See Time Averages (PASTA), the probability of finding the server occupied with a systematic packet is equal to the utilization factor of the server $\rho = \lambda T_{\text{tx}}$ [18]. Thus with probability ρ the second term of the delay due to redundancy is (approximately) given by Eqn. (3).

Let us now consider the case in which an arrival i arrives while the server is busy with a redundancy packet, which happens with probability $\rho\,(1/R-1)$ [3]. The second term of the delay due to redundancy is:

$$I_i^{\text{red}}(t) = T_{\text{tx}}\frac{N-K-1}{2}. \tag{4}$$

Plugging equations (3) and (4) into Eqn. (2), including the delay term $(N-K)\left\lfloor \frac{N_i(t)}{K} \right\rfloor$ due to redundancy and recalling that $X_j = T_{\text{tx}}$ we have :

$$\begin{aligned} W_i &\simeq S_i + T_{\text{tx}}\left[N_i(t) + (N-K)\left(\left\lfloor \frac{N_i(t)}{K} \right\rfloor + \right.\right. \\ &\left.\left. +\ \rho\frac{\text{mod}\ (N_i(t),K)+1}{K}\right) + \rho\left(\frac{1}{R}-1\right)\frac{N-K-1}{2}\right] \\ &\leq S_i + T_{\text{tx}}\left[N_i(t) + (N-K)\left(\frac{N_i(t)}{K} + \rho\frac{1}{K}\right)\right. \\ &\left. + \rho\left(\frac{1}{R}-1\right)\frac{N-K-1}{2}\right] \\ &= S_i + T_{\text{tx}}\left[N_i(t) + (N-K)\frac{N_i(t)}{K} + \rho\left(\frac{1}{R}-1\right)\frac{N-K+1}{2}\right]. \end{aligned} \tag{5}$$

Taking the limit for $i \to +\infty$ of the expectation at both sides of Eqn. (5) we get:

$$\begin{aligned} W &= \lim_{i\to+\infty} E\{W_i\} \\ &\simeq S_i + T_{\text{tx}}\left[N_Q + (N-K)\frac{N_Q}{K} + \rho\left(\frac{1}{R}-1\right)\frac{N-K+1}{2}\right], \end{aligned} \tag{6}$$

where $N_Q = \lim_{i\to+\infty} E\{N_i(t)\}$ is the average number of packets in the queue. Using the fact that, by Little's theorem, $N_Q = \lambda W$ [18], Eqn. (6) can be written as:

$$W \simeq \frac{S + \frac{T_{\text{tx}}\rho}{2}\left(\frac{1}{R}-1\right)(N-K+1)}{1-\frac{\rho}{R}}, \tag{7}$$

[2] This is in general not the case and is assumed here only to derive a closed form approximation.

[3] This can be easily proved taking into account that the ratio of time the server is occupied with redundancy packets to the time it is occupied with systematic packets is equal to $(N-K)/K = (1/R)-1$.

where $R = K/N$ is the rate of the network code. In order to calculate the residual service time S we use a procedure similar to that used in [21] to derive residual service time for a system with vacancy. Let us define $r(\tau)$ as the residual service time at time τ and let us fix an instant t in which the last redundancy packet of a block has just completed the service. The number of packets in the system in this type of instants constitutes an embedded Markov chain since the system is completely described by such number. The temporal average of the service time in $[0, t]$ is:

$$\begin{aligned}\frac{1}{t}\int_0^t r(\tau)d\tau &= \frac{1}{t}\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2 + \frac{1}{t}\sum_{i=1}^{V(t)}\frac{1}{2}X_i^2 \\ &= \left[\frac{M(t)}{t} + \frac{V(t)}{t}\right]\frac{T_{\text{tx}}^2}{2} \\ &= \left[\frac{M(t)}{t} + \frac{(N-K)}{K}\frac{M(t)}{t}\right]\frac{T_{\text{tx}}^2}{2}, \quad (8)\end{aligned}$$

where $M(t)$ and $V(t)$ are the number of systematic and redundancy packets transmitted in $[0, t]$, respectively. Note that, since in t the redundancy transmission for a block has just finished, $M(t)$ is an integer multiple of K while $V(t)$ is an integer multiple of $N - K$. Taking the limit for $t \to \infty$ at both sides of Eqn. (8) we obtain:

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t r(\tau)d\tau = \lim_{t\to\infty}\frac{M(t)}{t}\frac{N}{K}\frac{T_{\text{tx}}^2}{2} = \frac{\lambda}{R}\frac{T_{\text{tx}}^2}{2}. \quad (9)$$

The second equality in Eqn. (9) is valid in the stable region, since the average throughput $M(T)/t$ must be equal to the average arrival rate λ. If the time average can be replaced by the statistical average, we have $S = \frac{\lambda}{R}\frac{T_{\text{tx}}^2}{2}$ which, plugged into Eqn. (7), leads to the following approximation for the queueing delay at the transmitter side

$$W \simeq \frac{T_{\text{tx}}\rho}{2R}\frac{1 + (1-R)(N-K+1)}{1 - \frac{\rho}{R}}. \quad (10)$$

It is interesting to note that Eqn. (10) assumes the same expression of the delay in an M/D/1 queue if the rate R is set to one, which confirms the intuition that, if no redundancy is transmitted, the system behaves exactly as an M/D/1 queue. Note also that this is no longer valid if non-systematic network coding is used. It is also worth noting that, according to Eqn. (10), the (approximate formula for the) delay depends not only on the rate R, but also on the absolute number of redundancy packets transmitted per NC block $N - K$.

B. Receiver Side

The delay G at the receiver side includes the propagation delay and the decoding delay. The propagation delay is a fixed term equal to T_{pr} accounting for the time needed for each transmitted packet (systematic or redundancy) to reach the receiver. The decoding delay is the delay that affects a systematic packet in case it is not received correctly. If the packet is erased, the receiver will have to wait for enough redundancy to be received in order to recover the whole block. For the assumptions made on the code (large field size), the amount of redundancy needed to decode a lost packet is equal, with high probability, to the number of systematic packets that have been lost. Since the redundancy is transmitted after the systematic packets, the k-th packet of a block will have to wait for the arrival of the $K - k$ remaining packets before the redundancy starts to be transmitted. This implies that on average the decoding delay seen by a packet depends on the position of the packet in the block. We indicate the delay for packet in position k as $\overline{D}_k$. The average decoding delay is then:

$$\overline{D} = \frac{1}{K}\sum_{k=1}^{K}\overline{D}_k. \quad (11)$$

We model the channel between the transmitter and the receiver as an erasure channel with erasure probability $1 - p$. As usual practice in queue analysis, we assume the system is operated within the stable region, i.e., $N_Q < \infty$. In this case the arrival rate, modified to take into account the redundancy, λ/R, is equal to the rate packets are served by the transmitter, which coincides with the arrival rate at the receiver. However, the arrival process at the receiver is no longer Poisson, since the minimum interarrival time seen by the receiver is T_{tx}, which is the time needed by the transmitter to send a packet. In case of redundancy packets the interarrival time is exactly $T_{\text{sys}} = T_{\text{tx}}$, since redundancy packets are transmitted back-to-back. If the packets are systematic, on the other hand, the interarrival time depends on whether there are packets in the transmitter queue or not.

An upper bound on the average interarrival time between two systematic packets can be obtained by assuming that there is no packet waiting in the queue when the first packet is transmitted. In this case the interarrival time T_{sys} is the maximum between T_{tx} and a random variable exponentially distributed with mean value $\frac{1}{\lambda}$. The derivation is trivial and thus not reported here. The interarrival times for systematic and redundancy packets T_{sys} and T_{red} can thus be upper bounded as:

$$\begin{aligned}T_{\text{sys}} &\leq T_{\text{tx}}\left[\left(1 - e^{-\rho}\right) + e^{-2\rho} + \frac{e^{-2\rho}}{\rho}\right] \\ T_{\text{red}} &= T_{\text{tx}}.\end{aligned}$$

An upper bound on the average decoding delay for packet in position k is given by Eqn. (12). Note that in Eqn. (12) the j-th term in the first sum

$$\binom{K+j-2}{K-1}p^{K-1}(1-p)^{j-1}, \quad (13)$$

represents the probability to decode exactly $K - 1$ packets from the first $K - 1 + j - 1$ transmitted and is multiplied by p to take into account the fact that the additional delay after all systematic packets are transmitted is exactly $j \cdot T_{\text{tx}}$ if and only if the K-th correctly received packet is the j-th redundancy packet. Finally, the average delay at the receiver side is:

$$G_c = T_{\text{pr}} + \frac{1}{K}\sum_{k=1}^{K}\overline{D}_k. \quad (14)$$

C. PLR Analysis

Assuming an asymptotically large field size is used for the coefficients of the RLNC the PLR coincides with the probability to lose a systematic packet in the physical channel

$$\overline{D}_k \leq (1-p)\left[T_{\text{sys}}\cdot(K-k)+T_{\text{red}}\cdot\sum_{j=1}^{N-K} j\binom{K+j-2}{K-1}p^{K-1}(1-p)^{j-1}p+T_{\text{red}}\cdot(N-K)\sum_{t=0}^{K-1}\binom{N-1}{t}p^t(1-p)^{N-t-1}\right]. \quad (12)$$

conditioned to receiving less than K packets out of the remaning $N-1$, which is given by Eqn. (15):

$$\text{PLR} = (1-p)\sum_{j=0}^{K-1}\binom{N-1}{j}p^j(1-p)^{N-1-j}. \quad (15)$$

This result can be extended to the case of finite field size by applying the results in [22].

As a remark, we point out that the interdependency between the delays in equations (10) and (11) and the PLR in Eqn. (15) is hidden in the dependence on K, R and p.

IV. Numerical Results

With the results presented so far it is possible to determine an approximated region in the PLR-delay plane in which the system can operate. Note that the region we derive is an approximation due to the approximations used in the delay derivation. In the "achievable", by means of network coding we can reduce the PLR experienced by a QoS class at the price of an additional delay, caused by increased queueing delay and encoding/decoding in case of loss. In the following we present numerical results derived from the analysis carried out in Section III. Both the propagation delay and the transmission delay are set to 10 ms.

In Fig. 3 the analytical approximation for the queueing delay W is plotted against the system utilization factor ρ. In the same figure the actual queueing delay obtained through Monte Carlo simulation is also plotted. It can be seen how the approximation is pretty tight for low system loads, while it gets less tight as the load increases. This is due to the fact that, as it can be seen from Eqn. (4), ρ multiplies the term that has been approximated by excess, which implies that the difference with the actual value of the delay increases with ρ. Note that, although in this specific case the approximated expression leads to an upper bound, this is not in general the case. In the same figure the delay of an M/D/1 queue is also plotted. The arrival rate of the M/D/1, indicated as λ^{eq} in the figure, has been increased to λ/R (rather than λ) in order to make a fair comparison with the network-coded system. It is interesting to note how, even though the average number of packets processed by the network-coded system and the M/D/1 is the same, the delay in the NC system is much larger. This is due to the fact that the redundancy is transmitted back to back, keeping the server busy and increasing the waiting time in the queue, while in the M/D/1, since the incoming traffic is purely Poisson, packet arrivals are spread more evenly across time. Such difference also indicates that approximating (more precisely, bounding from below) the delay in a system with systematic NC using an M/D/1 model may not be accurate depending on the degree of precision sought.

In Fig. 4 we show the derived approximation on the total delay D_{tot} plotted against ρ/R for two different packet loss rates on the channel, namely $\varepsilon = 10^{-1}$ and $\varepsilon = 10^{-3}$. It is interesting to see how the total delay diverges as $\rho \to R$. This

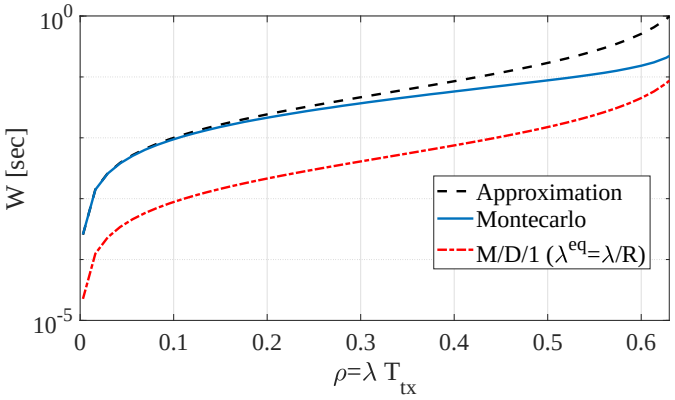

Fig. 3. Queueing delay W plotted against the system utilization factor ρ. The curve of the analytical approximation, the Monte Carlo simulation and the delay of an M/D/1 queue are shown. The arrival rate of the M/D/1, indicated as λ^{eq} in the figure, has been increased to λ/R in order to compensate for the redundancy in the network-coded system. $T_{\text{tx}} = 0.01$ seconds, $K = 60$ and $R = 2/3$ were used.

is due to the fact that the expression of the queueing delay, given by Eqn. (10), has a pole in $\rho/R = 1$. At low ε the delay is mainly due to the time spent in the queue, since little losses are experienced on the channel and thus the receiver does not need to wait for the redundancy packets (which, we recall, is due to the systematicity of the code). If ε is high, instead, there is an increase in the delay due to the losses, which forces the receiver to wait for the non systematic packets transmitted at the end of the block. The situation is exacerbated at low loads, as the time needed to receive a whole block is longer due to the relatively large interarrival time of packets at the receiver.

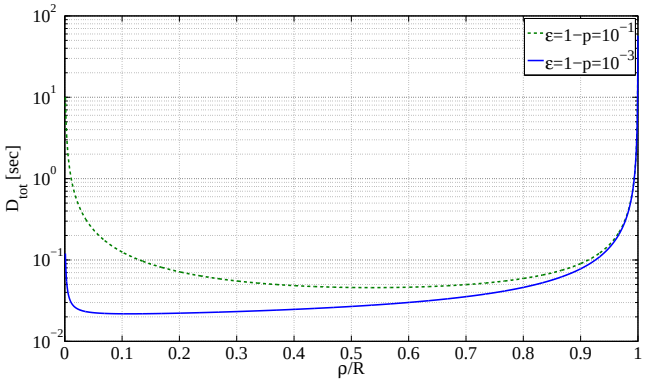

Fig. 4. Total delay (approximation) plotted against ρ/R for two different packet loss rates on the channel ε, namely $\varepsilon = 10^{-1}$ and $\varepsilon = 10^{-3}$. The plot was obtained fixing $K = 20$, $R = 0.9$ and varying the load λ.

In Fig. 5 the total delay D_{tot} is plotted against the PLR. The plot was obtained fixing $K = 100$ and varying the rate R in the range $[0.7, 1]$; the packet loss rate at the physical layer has been set to 20%. For larger values of R the PLR diminishes while the delay increases; the area to the right of the

dotted curve represents the region that is "achievable" through network coding. The packet loss rate at the physical layer ($\varepsilon = 1 - p$) is also shown with a thick vertical red line; a system without packet level coding would be forced to operate on such line. In order to see the impact of the block size K, in Fig.

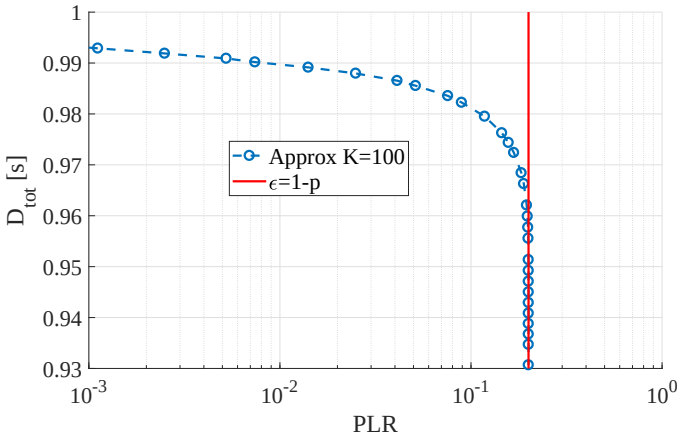

Fig. 5. Delay plotted against the packet loss rate. The plot was obtained fixing $K = 100$ and varying the rate R in the range $[0.7, 1]$. The packet loss rate at the physical layer has been set to 20% and is also shown in the figure. For larger values of R the PLR diminishes while the delay increases.

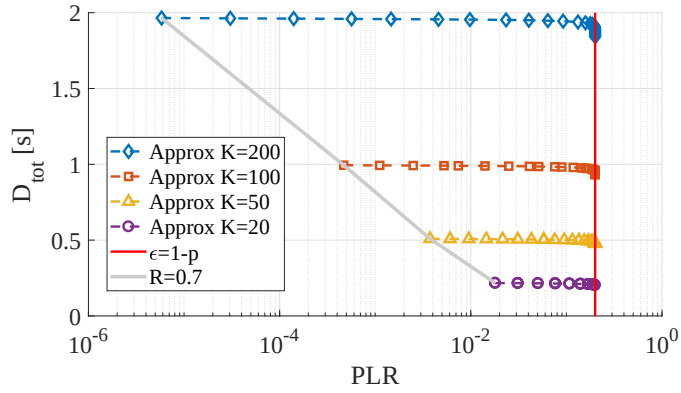

Fig. 6. Delay vs packet loss rate for different values of K. The plot was obtained by varying the code rate R in the range $[0.7, 1]$. The continuous grey line is the curve at constant $R = 0.7$. The packet loss rate on the channel $\varepsilon = 1 - p$ was set to 20%.

6 the delay vs packet loss rate for four different values of K is shown. The plot was obtained by varying the coding rate in the range $[0.7, 1]$ for all values of K. For ease of reading we plot the curve (continuous grey line) that joins all points at rate $R = 0.7$. It can be seen that the delay increases with K. This is due to the losses over the channel that force the receiver to wait for a whole block to decode a lost packet. On the other hand, it can be seen how the longer block length allows to achieve much better performance in terms of PLR as far as the code rate R is low enough. Finally, note that the PLR used in the figure is quite high (20%), which justifies the relatively high delays observed. The expressions we derived in the previous section can be used as a starting point to tune the code rate at the transmitter in order to find a good trade off between delay and PLR according to the desired QoS requirements.

V. Conclusions

We derived an approximated expression for the per-packet delay of a queueing system operated with systematic network coding and studied the trade-off between coding gain and decoding delay for finite-length blocks. A closed form expression has been derived for the approximation of the queueing delay which gives insight on the effect of important system parameters on the queueing delay. The approximation is quite tight at low system loads for the considered setups. The decreased PLR is provided at the cost of a slightly higher delay. In this respect we derived an approximated "achievable" region in the PLR-delay plane; this gives hints on whether the QoS requirements of a specific class of traffic can be fulfilled by applying systematic network coding. Currently we are working towards the extension of the work to the case of multiple classes of services and multiple links, as well as on the optimization of the system parameters.

Acknowledgment

This work has been partly supported by the European Commission FP7 Programme, in the BATS project, contract n.317533.

This is a revised version of the paper: G. Cocco, T. de Cola and M. Berioli, "Performance analysis of queueing systems with systematic packet-level coding", IEEE Int'l Conf. on Commun. (ICC), London,U.K., 2015.

This version first appeared on ResearchGate in October 2017.

References

[1] European Telecommunications Standards Institute, "ETSI TS 102 472 V1.1.1, Digital Video Broadcasting (DVB); IP Datacast over DVB-H: Content Delivery Protocols," June 2006.

[2] R. Bassoli, H. Marques, J. Rodriguez, K.W. Shum, and R. Tafazolli, "Network coding theory: A survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 1950–1978, Apr. 2013.

[3] J.K. Sundararajan, D. Shah, M. Medard, M. Mitzenmacher, and J. Barros, "Network coding meets tcp," in *IEEE INFOCOM 2009*, Apr. 2009, pp. 280–288.

[4] J.K. Sundararajan, D. Shah, and M. Medard, "ARQ for network coding," in *IEEE Int'l Symposium on Info. Theory (ISIT)*, July 2008.

[5] A. Eryilmaz, A. Ozdaglar, M. Medard, and E. Ahmed, "On the delay and throughput gains of coding in unreliable networks," *IEEE Trans. on Info. Theory*, vol. 54, no. 12, pp. 5511–5524, Dec. 2008.

[6] A.H. Salavati, B.H. Khalaj, P.M. Crespo, and M.-R. Aref, "QoSNC: A novel approach to QoS-based network coding for fixed networks," *Journal of Comm. and Networks*, vol. 12, no. 1, pp. 86–94, Feb. 2010.

[7] Y. Xuan and C.-T. Lea, "Network-coding multicast networks with QoS guarantees," *IEEE/ACM Trans. on Networking*, vol. 19, no. 1, pp. 265–274, Feb. 2011.

[8] W. Pu, C. Luo, F. Wu, and C.W. Chen, "QoS-driven network coded wireless multicast," *IEEE Trans. on Wireless Communications*, vol. 8, no. 11, pp. 5662–5670, November 2009.

[9] S. Bhadra and S. Shakkottai, "Buffer asymptotics for coding over networks," *IEEE Trans. on Info. Theory*, vol. 56, no. 12, pp. 6159–6181, Dec 2010.

[10] B. Shrader and A. Ephremides, "A queueing model for random linear coding," in *IEEE Military Communications Conference*, Orlando, FL, U.S.A., Oct. 2007.

[11] Y. Ma, W. Li, P. Fan, and X. Liu, "Queuing model and delay analysis on network coding," in *IEEE Int'l Symposium on Communications and Information Technology*, Beijing, China, Oct. 2005, vol. 1, pp. 112–115.

[12] B. Shrader and A. Ephremides, "On the queueing delay of a multicast erasure channel," in *IEEE Info. Theory Workshop (ITW)*, Punta del Este, Uruguay, Mar. 2006.

[13] B. Shrader and A. Ephremides, "Queueing delay analysis for multicast with random linear coding," *IEEE Trans. on Info. Theory*, vol. 58, no. 1, pp. 421–429, Jan. 2012.

[14] D.E Lucani, M. Medard, and M. Stojanovic, "Random linear network coding for time-division duplexing: Queueing analysis," in *IEEE Int'l Symposium on Info. Theory (ISIT).*, Seoul, Korea, June 2009.

[15] D.E. Lucani, M. Medard, and M. Stojanovic, "On coding for delay-network coding for time-division duplexing," *IEEE Trans. on Info. Theory*, vol. 58, no. 4, pp. 2330–2348, Apr. 2012.

[16] J. Barros, R.A. Costa, D. Munaretto, and J. Widmer, "Effective delay control in online network coding," in *IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009.

[17] D.E. Lucani, M. Medard, and M. Stojanovic, "Systematic network coding for time-division duplexing," in *IEEE Int'l Symposium on Info. Theory (ISIT)*, Austin, TX, U.S.A., June 2010, pp. 2403–2407.

[18] L. Kleinrock, *Queueing Systems*, vol. II, John Wiley and Sons, 1976.

[19] V.M. Vishnevsky, O.V. Semenova, A. Dudin, and V. Klimenok, "Queueing model with gated service and adaptive vacations," in *IEEE Int'l Conference on Communications Workshops (ICC Workshops)*, Dresden, Germany, June 2009, pp. 1–5.

[20] J. Chiarawongse, M.M. Srinivasan, and T.J. Teorey, "The M/G/1 queueing system with vacations and timer-controlled service," *IEEE Trans. on Communications*, vol. 42, no. 234, pp. 1846–1855, Feb. 1994.

[21] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Inc., first edition, 1987.

[22] R. Lidl and H. Niederreiter, *Finite Fields*, Cambridge University Press, 1997.