

Packet Skipping and Network Coding for Delay-Sensitive Network Communication

Marc Aoun^a, Paul Beekhuizen^b, Antonios Argyriou^c, Dee Denteneer^a, Peter van der Stok^a

^a*Philips Research, Eindhoven, 5656AE, The Netherlands*

^b*Eurandom, TU/e, Eindhoven, 5656AE, The Netherlands*

^c*Department of Computer and Communication Engineering, University of Thessaly, Volos, 38221, Greece*

Abstract

We provide an analytical study of the impact of packet skipping and opportunistic network coding on the timely communication of messages through a single network element. In a first step, we consider a single-server queueing system with Poisson arrivals, exponential service times, and a single buffer position. Packets arriving at a network node have a fixed deadline before which they should reach the destination. To preserve server capacity, we introduce a thresholding policy, based on remaining time until deadline expiration, to decide whether to serve a packet or skip its service. The obtained goodput improvement of the system is derived, as well as the operating conditions under which thresholding can enhance performance. Subsequently, we focus our analysis on a system that supports network coding instead of thresholding. We characterize the impact of network coding at a router node on the delivery of packets associated with deadlines. We model the router node as a queueing system where packets arrive from two independent Poisson flows and undergo opportunistic coding operations. We obtain an exact expression for the goodput of the system and study the achievable gain. Finally, we provide an analytical model that considers both network coding and packet skipping, capturing their joint performance. A comparative analysis between the aforementioned approaches is provided.

Keywords: Queueing model, network coding, packet skipping, real-time traffic, packet deadlines

1. Introduction

One way to interpret real-time communication is the assignment of particular importance to information flows in the time domain. In this case, the efficacy of data transfer is not measured in terms of raw communication throughput but it becomes dependent on the time instant the information is delivered. The increased interest in real-time communication comes at a time when the Internet becomes a carrier of more time-sensitive information and a new generation of systems such as wireless sensor networks are deployed for the monitoring of time-critical physical environments. High data rate applications like video-on-demand and high quality interactive video communications also require timeliness guarantees.

In this paper, we investigate the communication of real-time messages through a queueing theory -based analysis. We aim at modeling the achievable timeliness benefits when packet skipping and network coding are employed at a router node. We initially model a network node, where data packets have to be serviced (transmitted) within a certain deadline, as a single-server queue. The queueing system is assumed to have an overwrite-buffer with space for one packet; when a new packet bearing more recent data arrives to the queueing system, the buffered packet, if any, is discarded and the new packet is stored in the buffer. Including the buffer space at the server, the system represents an M/M/1/2 queue with deadlines.

The model we consider provides significant insights into the impact of local decisions on the timeliness performance of simple networks. In particular, we show that the real-time goodput of the queueing system is improved when the system employs a policy where buffered packets are only served if the remaining time until their deadline (i.e. their lead-time) is larger than a certain threshold θ . The intuition behind this policy is that server capacity is not wasted on packets that have a low probability of being served before their deadline expires. By skipping service of the currently stored packet, and waiting for a new packet arrival instead, goodput benefits are obtained because a newly arriving packet has a higher probability of meeting its deadline.

Under the assumption that the service times of the single-server system are exponentially distributed, we derive an exact analytic expression for the goodput of the system with this thresholding policy. The exact goodput expression allows us to compute the optimal threshold, which gives insight into the maximum performance gain, as well as the parameter ranges for

which the threshold policy provides effective performance benefits. In the second part of the paper, the M/M/1/2 queueing system is extended with the ability to algebraically code packets; with network coding, two data packets are transformed into one by a simple algebraic XOR-operation. The coded packet is transmitted and the original native packets can be retrieved at their respective destination nodes through a similar XOR-operation [1]. The idea of network coding has gained a lot of interest because of its potential to radically affect the way networks operate. When packets with timeliness restrictions undergo network coding, key QoS parameters might be improved but also compromised. Therefore, additional insight into the performance of systems with coding capability in the presence of timeliness requirements is needed. More specifically, it is critical to understand what is the impact of coding on packet delays and the interplay with the system goodput. In this part of the paper, we model a router node to which packets from two flows arrive and must be forwarded. We provide a detailed derivation and exact expressions for the stationary real-time goodput of the router node in the presence of network coding. The analytical results obtained in this paper can be related qualitatively to our recent simulation-based study of the network coding and thresholding policy [2]. Later, we will discuss aspects of these simulation results that can be related to this paper.

The rest of this paper is organized as follows. Section 2 provides background information on service models and the different types of deadlines that can be associated with real-time information. Related work is presented in Section 3. The system model adopted in this paper is presented in Section 4. The analysis for packet skipping based on lead-time thresholding is described in Section 5. Subsequently, the goodput analysis for a system with network-coded communication is developed in Section 6. Section 7 looks at the joint modeling of skipping and network coding. Finally, Section 8 presents our conclusions and possible directions for future work.

2. Service Models and Types of Deadlines

In queueing systems where packets are associated with deadlines, a service must usually either *begin* or *end* before a specific point in time. In the first case, services are usually non-preemptive (i.e., a service that has begun must be completed), while in the latter case, services are usually preemptive (i.e., a service that has begun can be aborted). In our case, the deadlines are to the end of service, but services are non-preemptive; indeed, the in-

formation contained in a packet needs to be completely transmitted before the receiver node can decode the packet and use the information, thus the end-of-service nature of the deadlines. Additionally, serving a packet in the context of wireless network communication consists of possibly contending for wireless channel access, and subsequent transmission of the packet bits. Due to the layered communication architecture adopted in wireless networks, little control can be exerted on the channel contention mechanism once it is activated. The same holds for the physical transmission once it is started.

Kruk et al. [3] furthermore consider three types of deadlines: Hard, firm, and soft. With hard deadlines, a single packet that is not served before its deadline makes all other packets useless as well. With this type of deadlines all packets must thus be served before their deadlines, which requires a worst-case analysis. With firm deadlines, every packet that is not served before its deadline is useless, whereas with soft deadlines lateness is permitted and packets can still be used afterwards. We consider the case of firm deadlines, a natural choice that arises when looking at back-end applications requirements in systems such as sensor networks, where conclusions should be based on non-expired data but deadline misses can be tolerated.

3. Related Work

Barrer [4, 5] was one of the first to analyze a queue with deadlines. He studied an $M/M/1$ queue with deterministic deadlines. This analysis was extended to systems with state-dependent arrival and service rates, and more general deadlines by Brandt and Brandt [6, 7] and by Movaghar [8]. These studies all deal with non-preemptive systems with deadlines to service beginnings, but in [9], Movaghar considers deadlines to service endings with preemption. Movaghar and Kargahi [10, 11] have devised an approximation for an $M/M/1$ queue with the earliest-deadline-first (EDF) discipline, which is known to stochastically maximise the fraction of packets served before their deadline (see, e.g., [12]).

Lehoczky [13] analyzed an $M/M/1$ queue with deadlines to service endings and the EDF-policy in heavy traffic. He argued that, since the deadlines of all stored packets have to be taken into account, this queue gives rise to a Markov process on a statespace of infinite dimension. Lehoczky shows that the Markov process collapses to a tractable one-dimensional process in heavy traffic. Lehoczky later used these results to analyze control policies in [14], and extended his analysis to Jackson networks in [15]. Doytchinov et al. [16]

extended this analysis to a $GI/G/1$ queue with deadlines, and Kruk et al. [3] to networks of such queues.

Delay sensitive traffic in the presence of network coding was studied by Parag et al. in [17]. The authors adopted a statistical QoS measure that expresses the decay rate of the buffer at the middle node in a butterfly network. This router node is the bottleneck in the butterfly case, which explains the reasons why the authors focus on its buffer behavior. Although this metric is enough for approximating the delivered QoS per flow, it does not express it directly in terms of the achieved packet error rate and goodput. Shah et al. [18] start with the goal of minimizing the backlog of coded packets at receiving nodes. They design an online algorithm so that the linear packet combinations that are generated, are chosen in such a way that their actual span excludes any linear combination that is already known to the receivers.

Eryilmaz et al. [19] studied the delay benefits of network coding in wireless multicast and multiple unicast scenarios. They presented a model that considers only a single-hop transmission and the random coding across packets from the same flow (intra-session network coding). Online network coding and delay minimization was more recently presented in [20]. A precise model is not presented in that work, although a simple delay analysis for the wireless channel with Bernoulli erasures is performed. Another interesting modeling work can be found in [21] where the authors used stochastic network calculus for calculating the throughput in a coded butterfly network. Nevertheless, the model does not cover delay and timeliness aspects. A recent work from Goseling et al. [22] aims at modeling the performance of coded queueing systems. Instead of focusing on providing exact expressions, the authors provide bounds on the performance of a coded tandem queueing network with two independent flows. Nevertheless, the extension of this work to packets with deadline requirements is not straightforward.

4. System Model

For the analysis of packet skipping, we consider the single server queueing system depicted in Fig. 1. Packets arrive according to a Poisson process with parameter λ and have exponentially distributed service times with parameter μ . The system has an ‘overwrite’ buffer with one buffer position (not counting the packet in service, if there is one). If a packet arrives when there is already a packet in the buffer, the latter is overwritten (i.e., replaced) by the newly arriving packet. The result is thus a modified version of an $M/M/1/2$

queueing system (note that, in an $M/M/1/K$ system, the K refers to the number of buffer positions, including the one in service).

The same $M/M/1/K$ system is adopted for the analysis of network coding; Two input processes with parameters λ_1 and λ_2 are considered, as depicted in Fig. 2. The overall arrival process here is also a Poisson process with parameter $\lambda = \lambda_1 + \lambda_2$. In the network coding case, when a packet is waiting in the buffer, a second packet arriving to the system is algebraically coded with the waiting packet instead of overwriting it, in case they belong to different flows. Overwriting takes place when waiting packet and arriving packet belong to same flow, or when the buffer is already occupied by a coded packet. With network coding the coded packet is transmitted and the original native packets can be retrieved at their respective destination nodes through a similar algebraic XOR operation [1]. We provide two scenarios where the above is possible. In the case of multicast transmission if each node has one of the two desired packets, they can both recover the other packet after the reception of the single XOR-coded packet [1]. Also a similar situation arises in the wireless bidirectional relay network where both nodes want to transmit to each other a packet they already have through an intermediate relay node [2]. In this case similarly, if the two nodes have packets a and b , with the transmission of a single coded packet $a \oplus b$ from the relay they could both decode the opposite packet. The broadcast nature of the wireless medium allows this operation.

A packet service is considered successful if it is *completed* within d time units after the arrival of the packet to the queue, i.e., if a packet arrives at time t , its service must end before its global deadline $t + d$, with d being a fixed relative deadline. Services are non-preemptive, which means that if a packet starts service, its service has to be completed, before service on another packet can be started. We define the goodput of the system as the number of packets that are served successfully, i.e. within the deadline requirement, per unit of time.

In the packet skipping case, we introduce a threshold θ such that the packet in the buffer may only start service if its lead-time, i.e. the remaining time until its absolute deadline, is larger than θ . If the time until the deadline of the packet becomes smaller than θ , the packet is removed from the buffer. The idea behind this threshold is that if the deadline of a packet will be exceeded shortly, there is a large probability that the packet is not served in time. Thus, the goodput of the system might be improved by waiting for another packet that has a higher probability of being served successfully.

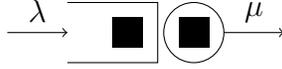


Figure 1: A schematic representation of the model for a node adopting packet skipping.

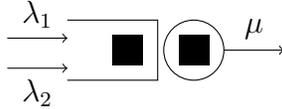


Figure 2: A schematic representation of the model with the router exercising network coding.

A packet whose deadline has expired is not served, both in the system that employs packet skipping and in the system that employs network coding. A coded packet is considered to be expired if the deadlines of both its underlying native packets are expired. This is implemented by setting the expiration deadline of a coded packet to that of the largest among the absolute deadlines of the underlying native packets that were coded together. Note that the individual deadlines of the native packets are still the ones considered when calculating the goodput. Therefore, a coded packet that completes service before its expiration deadline does not necessarily mean that the underlying native packet with the smallest deadline has been served on-time. Also a packet that arrives to an empty system is always served. Indeed, with fixed deadlines, for packet skipping, there is no gain in *not* serving packets arriving to an empty system and waiting for the next packet. After all, the next packet will have the same deadline, and hence the same probability of being served successfully. Delaying the service of a packet arriving to an empty system might be proved to provide additional benefits in the network coding case, but was not considered in this work due to the additional analysis complexity it would engender.

5. Performance Analysis of Packet Skipping

5.1. Goodput Analysis

In this section, we compute the goodput of the system in the presence of packet skipping. It will be convenient to view the goodput as a function of the threshold θ , and we denote it by $\gamma(\theta)$. Furthermore, we denote by L the

number of packets present in the system (including the server) immediately before a packet arrival.

A packet arriving to the system will find the latter in one of the following three states:

- *Empty* ($L = 0$): The server is available, and there is no packet occupying the waiting buffer.
- *Busy Server* ($L = 1$): The server is servicing a packet, but the waiting buffer is free.
- *Full*, ($L = 2$): the server is busy and the waiting buffer is occupied. In that case, the arriving packet overwrites the buffered one.

Theorem 1. *The goodput $\gamma(\theta)$ is given by*

$$\begin{aligned} \gamma(\theta) = & \lambda \mathbb{P}(L = 0)(1 - e^{-\mu d}) \\ & + \lambda (1 - \mathbb{P}(L = 0)) \left[\frac{\mu}{\lambda + \mu} (1 - e^{-(\lambda + \mu)(d - \theta)}) \right. \\ & \left. - \frac{\mu e^{-\mu d}}{\lambda} (1 - e^{-\lambda(d - \theta)}) \right], \end{aligned} \quad (1)$$

where

$$\mathbb{P}(L = 0) = 1 - \frac{\lambda^2 + \lambda\mu}{\mu^2 + \lambda\mu e^{-(\lambda + \mu)(d - \theta)} + \lambda^2 + \lambda\mu} \quad (2)$$

is the probability that an arbitrary packet arrives at an empty system.

Proof. The goodput $\gamma(\theta)$ is by definition equal to the arrival rate times the probability that the service of an arbitrary packet is successful. To compute this probability, we condition on whether a packet arrives at an empty system or not. Denoting by B the service time that the arriving packet would experience, we have:

$$\begin{aligned} \gamma(\theta) = & \lambda \mathbb{P}(L = 0) \mathbb{P}(B \leq d) \\ & + \lambda (1 - \mathbb{P}(L = 0)) \int_0^{d - \theta} \mu e^{-\mu t} \mathbb{P}(B \leq d - t) e^{-\lambda t} dt \end{aligned} \quad (3)$$

Evaluating the probabilities and the integral indeed yields (1).

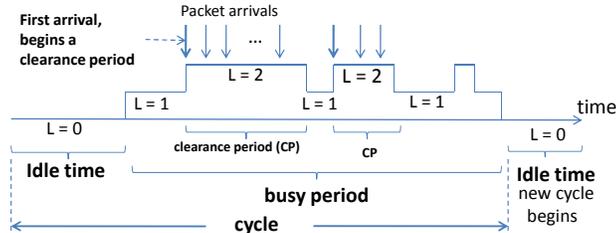


Figure 3: A cycle consisting of idle time and busy period.

The rationale behind Eq. (3) is as follows: If an arbitrary packet P arrives at an *empty* system, its service is successful if its service time B is less than d . This explains the term $\mathbb{P}(L = 0)\mathbb{P}(B \leq d)$. If P arrives at a *non-empty* system ($L = 1$ or $L = 2$), we condition on the length of the remaining (residual) service time t of the packet in service. The residual service time is exponentially distributed, so its density is $\mu e^{-\mu t}$. Moreover, given that the residual service time is t , the service of P is successful if the following three conditions are all met: First, the time until the deadline is larger than θ when the residual service ends, i.e., $d - t \geq \theta$. This is taken into account in the integration region of Eq. (3). Second, the service time is less than $t - d$, which explains the factor $\mathbb{P}(B \leq t - d)$. Third, there were no other arrivals during t time units, which explains the term $e^{-\lambda t}$.

In order to determine $\mathbb{P}(L = 0)$, we introduce the concept of cycle time, which we define as the time between two consecutive moments at which the system becomes empty. The cycle time C consists of two parts: A part during which the system is empty, the ‘idle time’, and a part during which the system serves packets until it becomes empty again, the ‘busy period’. An illustration of a cycle is presented in Fig. 3.

The busy period spreads throughout the time that the system is not in a state $L = 0$. It consists of a collection of consecutive states characterized by $L = 1$ or $L = 2$. We shall refer to the case where $L = 2$ during a busy period as *clearance period (CP)*, which is the amount of time the buffer is continuously occupied. A clearance period ends when the buffer is cleared, i.e. when either a dispatch to the server occurs, or when the stored packet is denied service and removed from the system following the thresholding procedure.

The idle time lasts until a packet arrives, so the mean idle time is $1/\lambda$. Denoting the busy period by BP , we have:

$$\mathbb{E}[C] = \frac{1}{\lambda} + \mathbb{E}[BP].$$

The PASTA-property (Poisson Arrivals See Time Averages, see [23]) implies that the probability that a packet arrives at an empty system, $P(L = 0)$, is given by the probability that the system is empty at an arbitrary time. A standard argument from renewal theory implies that the latter probability is given by the mean idle time divided by the mean cycle time:

$$\mathbb{P}(L = 0) = \frac{1/\lambda}{\mathbb{E}[C]} = \frac{1/\lambda}{1/\lambda + \mathbb{E}[BP]}. \quad (4)$$

The mean busy period is given by the following equation:

$$\mathbb{E}[BP] = \frac{1}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} \cdot 0 + \frac{\lambda}{\lambda + \mu} (\mathbb{E}[T] + \mathbb{E}[BP]), \quad (5)$$

where $\mathbb{E}[T]$ is the mean buffer clearance period. We will derive $\mathbb{E}[T]$ next.

Eq. (5) follows from a probabilistic argument: The busy period starts when a packet arrives to an empty system. After this, the expected time until the first event (an arrival or service completion) is $1/(\lambda + \mu)$. Furthermore, with probability $\mu/(\lambda + \mu)$, the first event is a service completion, and the busy period ends after 0 additional time units. With probability $\lambda/(\lambda + \mu)$, the first event is an arrival. In this case, the arriving packet is stored in the buffer and the buffer has to be cleared, which takes in expectation an additional $\mathbb{E}[T]$ time units. Once the buffer has been cleared, the expected time before the system becomes empty is again equal to $\mathbb{E}[BP]$ due to the memorylessness of service times. Evaluating (5) yields

$$\mathbb{E}[BP] = \frac{1}{\mu} + \frac{\lambda}{\mu} \mathbb{E}[T]. \quad (6)$$

It thus remains to determine $\mathbb{E}[T]$, the expected length of the buffer clearance period. Without loss of generality, we assume that the buffer clearance period starts at time 0. We condition on t , the time at which the next event (arrival or service completion) occurs. Because the time until the next event

is exponentially distributed with parameter $\lambda + \mu$, we obtain:

$$\begin{aligned}\mathbb{E}[T] &= \int_{d-\theta}^{\infty} (\lambda + \mu)e^{-(\lambda+\mu)t}(d - \theta)dt \\ &+ \int_0^{d-\theta} \frac{\mu}{\lambda + \mu}(\lambda + \mu)e^{-(\lambda+\mu)t}tdt \\ &+ \int_0^{d-\theta} \frac{\lambda}{\lambda + \mu}(\lambda + \mu)e^{-(\lambda+\mu)t}(t + \mathbb{E}[T])dt.\end{aligned}$$

The three integrals (from left to right) cover the following three possibilities: First, if there are no events before time $d - \theta$, the buffer is cleared at time $d - \theta$ because the deadline of the buffered packet becomes smaller than θ . Second, if the first event occurs at time $t < d - \theta$, and the first event is a service completion, the buffer is cleared at time t because the buffered packet enters service. Third, if the first event occurs at time $t < d - \theta$ and is a packet arrival, the arriving packet overwrites the buffered packet, and a new buffer clearance period begins.

After rewriting and evaluating the integrals (the rightmost two using partial integration), we obtain:

$$\mathbb{E}[T] = \frac{1 - e^{-(\lambda+\mu)(d-\theta)}}{\mu + \lambda e^{-(\lambda+\mu)(d-\theta)}}. \quad (7)$$

Combining Eqs. (4), (6), and (7) yields Eq. (2). \square

Remark 1. If $d \rightarrow \infty$, deadlines become irrelevant and two special cases occur for specific values of θ . If $\theta = 0$, packets in the buffer are always served, regardless of the time until their deadline. The goodput of the system is thus equal to that in an $M/M/1/2$ queue, namely $\lambda(1 - \frac{1}{1+\rho+\rho^2})$, with $\rho = \lambda/\mu$ (see e.g., [24, Section 5.7] or [25, Section 3.6]). If $\theta = d \rightarrow \infty$, packets in the buffer are never served. Only packets arriving to an empty system are. In this case, the goodput is equal to that in an $M/M/1/1$ queue, namely $\lambda(1 - \frac{1}{1+\rho})$. By substituting d and θ , these values follow from Eq. (1) as well.

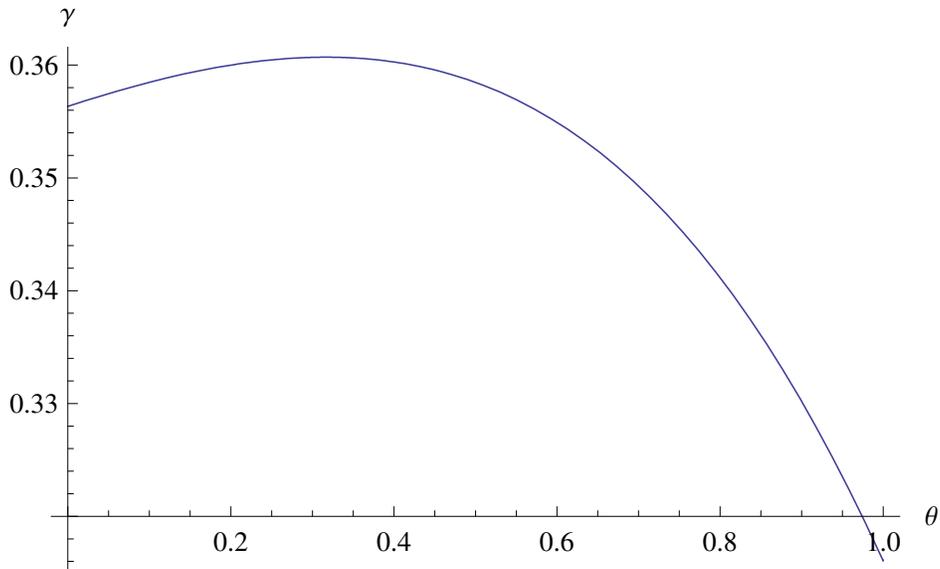


Figure 4: The goodput as a function of θ , with $\mu = \lambda = d = 1$.

5.2. Numerical results

In this section, we study the goodput numerically. In Fig. 4, we present $\gamma(\theta)$ for various values of θ , with $\mu = \lambda = d = 1$. We clearly see that the goodput of the system is indeed increased by the threshold θ , as long as θ is chosen appropriately.

Having established that an appropriately chosen threshold increases goodput, we study *how much* the goodput can be increased by this threshold. Note that $\gamma(0)$ is the goodput of the system without threshold. We define θ^* as the threshold that maximizes the relative increase in goodput, i.e., we define

$$\theta^* = \arg \max \left\{ \frac{\gamma(\theta) - \gamma(0)}{\gamma(0)} : 0 \leq \theta \leq d \right\}.$$

We define the maximal goodput gain as the maximal relative increase in goodput, i.e., as $(\gamma(\theta^*) - \gamma(0))/\gamma(0)$.

In the sequel, we fix $\mu = 1$. We can make this assumption without loss of generality; all parameters are relative to each other, so for any set of parameters, we can scale time in such a way that $\mu = 1$ without changing the goodput of the system.

In Fig. 5 we display the maximal goodput gain for d between 0 and 2, and

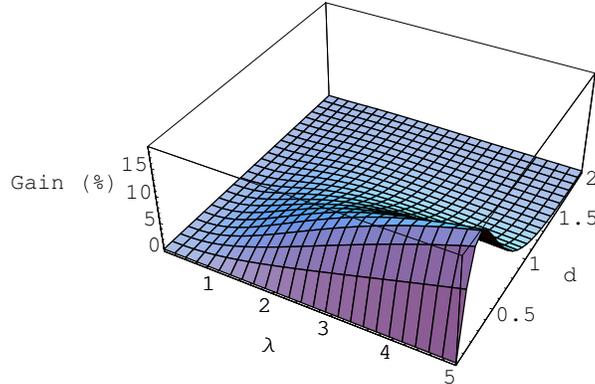


Figure 5: The maximal goodput gain for $0 \leq \lambda \leq 5$.

λ between 0 and 5. Fig 5 implies that, in this parameter region, we can obtain an increase of up to 15% in goodput by setting the threshold to its optimal value. Furthermore, we see that the relative increase in goodput is maximal if $d \approx 0.3$, but even for larger values of d there can be a goodput enhancement. In addition to this, the maximal goodput gain grows as λ grows, so the goodput policy is especially beneficial if the system is overloaded.

Likewise, in Fig. 6 we show the maximal goodput gain for λ up to 10. In this region, the increase in goodput can be as much as 30%. Furthermore, the value of d with the largest relative increase in goodput, as well as the range of d for which gain is achieved become smaller. This is conveyed by Fig. 7, which plots the gain for three increasing values of λ , equal to 3, 4 and 6. The narrower range of d where gain is achieved is explained by a shorter sojourn time of buffered packets that later enter service, due to more frequent buffer overwriting for an increasing arrival rate. The shorter sojourn time reflects itself in a larger lead-time when the buffered packet is assessed, by the packet skipping component, for dispatch to the server. This in turns presents itself in a decrease of the upper value of d for which packet skipping still results in service denials for some packets, and as such, still results in goodput gain.

Finally, we show values of the gain for specific values of λ and d in Table 1. We clearly see that, for this parameter region, the maximal goodput increase grows as λ becomes larger. Furthermore, for a fixed λ , the maximal gain

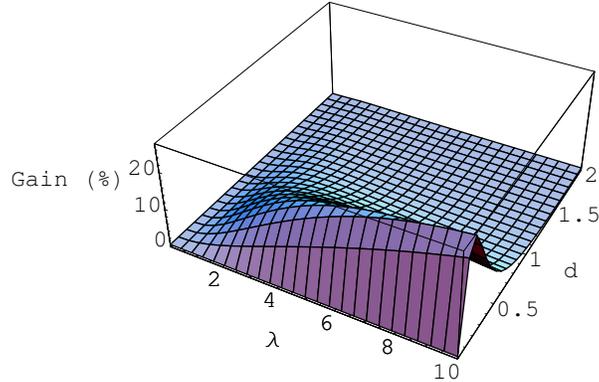


Figure 6: The maximal goodput gain for $0 \leq \lambda \leq 10$.

Gain (%)	d				
	0.1	0.2	0.3	0.4	0.5
0.5	0.23	0.37	0.45	0.48	0.48
1	1.04	1.71	2.07	2.21	2.18
λ 2	1.036	5.90	6.91	7.05	6.64
4	10.07	14.69	15.30	13.78	11.49
8	21.36	24.20	19.61	14.31	10.08

Table 1: The maximal goodput gain for specific values of λ and d .

increases up to a certain value of d , and decreases beyond that value.

6. Performance Analysis of Network Coding

In this part of the paper we focus our analysis on a router element R, to which two flows of packets arrive and need to be routed. The system model is depicted in Fig. 2. The router node is capable of opportunistically coding two native (un-coded) packets belonging to different flows, and transmitting the resulting coded packet instead of two native transmissions. For convenience, we refer to packets of the first flow as type-1 packets, while those of the second flow are type-2 packets. A coded packet is referred to as type-3. The sum of the two arrival rates λ_1 and λ_2 is denoted by λ .

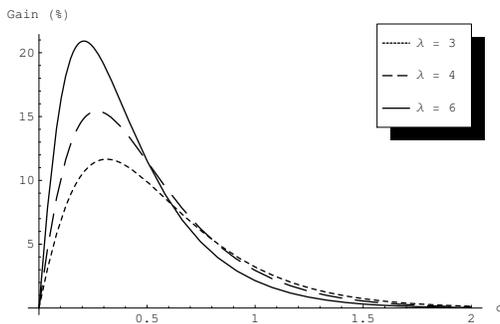


Figure 7: Goodput gain for λ values equal to 3, 4 and 6.

We shall restrict the analysis to one relative deadline d value for both flows. This restriction neither impacts the quality nor the representativeness of the analysis. It is merely used to simplify the underlying presentation. The extension of our analysis to different deadline distribution is straightforward.

6.1. Goodput Analysis

We define the *goodput of a flow* as the number of packets of that flow that are served by the router within their deadline requirement, per unit of time. Given the symmetry of the problem, we will concentrate our analysis on the goodput of type-1 packets. The analysis for packets of type-2 is identical.

The state of the queueing system at the router is fully conveyed through the characterization of the status of its waiting buffer and the status of its server. Two variables will be used to keep track of the system state: L , already introduced in Section 5.1, that defines the current number of packets in the system (including a packet being served), and N , that defines the type of the packet residing in the buffer, if any. Knowledge of the type of a packet when it is in service is not required for the analysis.

Variable L can take one of three values: 0, 1, or 2. Note that a coded packet is counted as a single packet in the system. Variable N can take one of three values: 1 (packet in buffer is of type-1), 2 (packet in buffer is of type-2), or 3 (packet in buffer is a coded packet, type-3).

A packet of type-1 arriving to the system can find it in one of five states:

- *Empty* ($L = 0$): The server is available, and there is no packet occupying the waiting buffer.

- *Busy Server* ($L = 1$): The server is servicing a packet, but the waiting buffer is free.
- *Full, type-1 packet in buffer* ($L = 2, N = 1$): the server is busy. The waiting buffer is occupied by a type-1 packet.
- *Full, type-2 packet in buffer* ($L = 2, N = 2$): the server is busy. The waiting buffer is occupied by a type-2 packet.
- *Full, type-3 packet in buffer* ($L = 2, N = 3$): the server is busy. The waiting buffer is occupied by a coded packet.

Theorem 2. *The goodput of type-1 packets is given by the following expression:*

$$\gamma_1 = \lambda_1 \sum_{a,b} \mathbb{P}(L = a, N = b) \mathbb{P}(\text{success} | L = a, N = b), \quad (8)$$

where $\mathbb{P}(L = a, N = b)$ is the probability that the system is in state ($L = a, N = b$) and $\mathbb{P}(\text{success} | L = a, N = b)$ is the probability that an arbitrary type-1 packet arriving to the system in state ($L = a, N = b$) meets its deadline.

Proof. Follows a similar reasoning as the proof of Theorem 1; The goodput γ_1 of type-1 packets is by definition equal to the probability that the service of an arbitrary packet of flow 1 is successful, i.e. meets its deadline. By the PASTA property [23], the probability that an arbitrary type-1 packet arrives to the system in state ($L = a, N = b$) is equal to the steady-state probability that the system is in state ($L = a, N = b$). The overall service success probability is therefore given by summing, over the entire state space, the probability that the system is in state ($L = a, N = b$) times the success probability conditioned on the occurrence of this state. \square

Finding γ_1 amounts to finding the success probabilities per system state, $\mathbb{P}(\text{success} | L = a, N = b)$, and the different state probabilities $\mathbb{P}(L = a, N = b)$. We derive these probabilities in the following two subsections.

Success Probability per System State

We start by finding the probabilities $\mathbb{P}(\text{success} | L = a, N = b)$ that a packet of type-1 arriving to the system when it is in state ($L = a, N = b$) is served by the router server and completes its service before the expiration of its deadline. The arrival of the packet is considered as time origin. Therefore, the absolute deadline of the packet is equal to its relative deadline.

Case 1: Arrival to an Empty System ($L = 0$)

A packet arriving to an empty system directly enters service. Its success probability is equal to:

$$\mathbb{P}(\text{success}|L = 0) = (1 - e^{-\mu d}), \quad (9)$$

which is equal to the probability that the service time it experiences is smaller than its deadline d .

Case 2: Arrival to a Busy Server ($L = 1$) or Full System ($L = 2$) with $N = 1$ or 3

Due to the overwrite property of the waiting buffer, an arrival of a type-1 packet to a Busy Server state is identical to an arrival when the system is full and the buffered packet is of type-1 or type-3. Indeed, in all three cases, the newly arriving packet will occupy (or overwrite) the buffer and wait to be serviced. Therefore $\mathbb{P}(\text{success}|L = 2, N = 1)$ and $\mathbb{P}(\text{success}|L = 2, N = 3)$ are both equal to $\mathbb{P}(\text{success}|L = 1)$, given by:

$$\begin{aligned} \mathbb{P}(\text{success}|L = 1) &= \int_0^d \mu e^{-\mu t} e^{-\lambda_1 t} (1 + \lambda_2 t) e^{-\lambda_2 t} \\ &\quad \times (1 - e^{-\mu(d-t)}) dt. \end{aligned} \quad (10)$$

To derive Eq. (10), we first condition on the length of the residual service time t of the current packet in the server. Being exponentially distributed, this residual service time has a density equal to $\mu e^{-\mu t}$. Given that the residual service time is t , a buffered packet is successfully serviced if all following conditions are met:

1. The server becomes free before the deadline expiration of the buffered packet, i.e., $t \leq d$.
2. The packet is not overwritten while waiting in the buffer.
3. The service time of the packet is smaller than the remaining time the packet has until deadline expiration.

Condition 1 is accounted for in the integration region (0 to d). Condition 2 is met if and only if, starting from the arrival moment 0 of the buffered packet until the residual service time t is completed, no new type-1 packet

arrives and at most one arrival of type-2 occurs. Indeed, the first arrival of a type-2 packet will not result in overwriting the buffered packet, since it will be coded with it (resulting in a type-3 packet). Any arrival of type-1 is not tolerated since it overwrites both a type-1 and a type-3 buffered packet. The term $e^{-\lambda_1 t}$ gives the probability that no type-1 arrivals occur during t . The term $(1 + \lambda_2 t)e^{-\lambda_2 t}$ gives the probability that at most one type-2 arrival occurs.

Finally, condition 3 is met if the service time experienced by the packet is *at most* equal to the remaining time $(d - t)$ until deadline expiration. The term $(1 - e^{-\mu(d-t)})$ is equal to the probability that the service time is smaller than $d - t$.

6.1.1. Case 3: Arrival to a Full System with Buffered Type-2 Packet ($L = 2$, $N = 2$)

If the type-1 packet arrives to a system where the waiting buffer is occupied by a type-2 packet, both packets will be coded together, resulting in a type-3 buffered packet. The success probability of the arriving packet is therefore equal to the probability that the resulting type-3 packet is served before the deadline d of the arriving packet is expired. It is given by:

$$\mathbb{P}(\text{success} | L = 2, N = 2) = \int_0^d \mu e^{-\mu t} e^{-\lambda_1 t} e^{-\lambda_2 t} \times (1 - e^{-\mu(d-t)}) dt. \quad (11)$$

Eq. (11) differs from Eq. (10) in that no arrival of type-2 is allowed at all, throughout the whole sojourn time of the coded packet in the buffer. Indeed, any type-2 arrival will overwrite the coded packet. This arrival restriction on type-2 packets is reflected in Eq. (11) through the term $e^{-\lambda_2 t}$.

System State Probabilities

In order to determine the probabilities $\mathbb{P}(L = a, N = b)$ of finding the system in a particular state $(L = a, N = b)$, we again make use of the notion of cycle time. For additional clarity of the following analysis, the reader is referred to Fig. 3.

The arrival that causes a state change from $L = 1$ to $L = 2$ (i.e. that starts a clearance period) is referred to as *first arrival*. As a busy period

might comprise multiple clearance periods, there can be multiple first arrivals during a busy period. Note that the notion of a type-3 packet as first arrival is abstract in the sense that it does not occur in practice, but is introduced for analysis purposes.

The idle time lasts from the moment the system becomes empty until the arrival of a packet of either type-1 or type-2. Therefore, the mean idle time is equal to $1/(\lambda_1 + \lambda_2)$. The mean busy period is the same as that given in Eq. (6), with λ equal to the sum of λ_1 and λ_2 . Similarly, the expected length of the buffer clearance period, $\mathbb{E}[T]$, is obtained from Eq. (7) by setting θ to 0. The mean cycle time $\mathbb{E}[C]$ is obtained as the sum of the mean idle time and the mean busy period.

As standard argument from renewal theory, the probability that the system is in a certain state is given by the mean time the system spends in this state during a cycle, divided by the mean cycle time. Let C_i be the total time the system has a packet of type i in the buffer during a cycle time. In other words, C_i is the total amount of time the system is in state ($L = 2, N = i$) during a cycle. The system state probabilities are then obtained using:

$$\mathbb{P}(L = 0) = \frac{\mathbb{E}[I]}{\mathbb{E}[C]}, \quad (12)$$

$$\mathbb{P}(L = 2, N = i) = \frac{\mathbb{E}[C_i]}{\mathbb{E}[C]}, \text{ and} \quad (13)$$

$$\mathbb{P}(L = 1) = 1 - P(L = 0) - \sum_{i=1}^3 P(L = 2, N = i). \quad (14)$$

To find $\mathbb{E}[C_i]$, we define $T_{i,j}$ as the cumulative amount of time that the buffer is occupied by type- i packets during a clearance period, given that the *first arrival* of that clearance period was a type- j packet. The expected value of C_i is given by simply summing up the two different cases of packet types for the first arrival:

$$\begin{aligned} \mathbb{E}[C_i] &= \sum_{j=1}^2 \frac{\lambda_j}{\lambda + \mu} (\mathbb{E}[T_{i,j}] + \mathbb{E}[C_i]) \\ &= \sum_{j=1}^2 \frac{\lambda_j}{\mu} (\mathbb{E}[T_{i,j}]). \end{aligned} \quad (15)$$

It now finally remains to find the expected values of the different $T_{i,j}$. The expected value of $T_{1,1}$ is given by:

$$\begin{aligned} \mathbb{E}[T_{1,1}] = \int_0^d (\lambda + \mu)e^{-(\lambda+\mu)t} & \left[\frac{\mu}{\lambda + \mu}t + \frac{\lambda_1}{\lambda + \mu} (E[T_{1,1}] + t) \right. \\ & \left. + \frac{\lambda_2}{\lambda + \mu} (E[T_{1,3}] + t) \right] dt + de^{-(\lambda+\mu)d}. \end{aligned} \quad (16)$$

Eq. (16) conveys the following: After the *first arrival* an event is bound to occur: either a service completion, or a new arrival. We condition on t , the time at which the next event occurs. Because of the memorylessness of the system, t has density $(\lambda + \mu)e^{-(\lambda+\mu)t}$. With probability $\frac{\mu}{\lambda+\mu}$, this event is a service completion, and the clearance period is equal to t . With probability $\frac{\lambda_2}{\lambda+\mu}$, the first event is a type-2 arrival. Due to memorylessness, this can be seen as the beginning of a new clearance period with a type-3 arrival as *first arrival*, i.e., with expectation $\mathbb{E}[T_{1,3}]$. However, up to time t , there was a type-1 packet in the buffer, so the expected *cumulative* amount of time type-1 packets spend in the buffer is given by $t + \mathbb{E}[T_{1,3}]$. With probability $\frac{\lambda_1}{\lambda+\mu}$, the first event is a type-1 arrival. Likewise, this can be seen as a new clearance period with a type-1 arrival as a *first arrival*, i.e., with expectation $\mathbb{E}[T_{1,1}]$ and cumulative amount of time equal to $t + \mathbb{E}[T_{1,1}]$. Finally, with probability $e^{-(\lambda+\mu)d}$, no event happens prior to the deadline expiration and subsequent removal of the type-1 first arrival. In that case, the buffer would have been occupied by this type-1 packet for an amount of time equal to d .

Likewise, the expected values of $T_{1,2}$ and that of $T_{1,3}$ are respectively given by:

$$\begin{aligned} \mathbb{E}[T_{1,2}] = \int_0^d (\lambda + \mu)e^{-(\lambda+\mu)t} & \\ & \left[\frac{\lambda_1}{\lambda + \mu} E[T_{1,3}] + \frac{\lambda_2}{\lambda + \mu} E[T_{1,2}] \right] dt, \end{aligned} \quad (17)$$

$$\mathbb{E}[T_{1,3}] = \int_0^d (\lambda + \mu) e^{-(\lambda+\mu)t} \left[\frac{\lambda_1}{\lambda + \mu} E[T_{1,1}] + \frac{\lambda_2}{\lambda + \mu} E[T_{1,2}] \right] dt. \quad (18)$$

Let $\alpha = 1 - e^{-(\lambda+\mu)d}$. The upper set of integrals results in the following set of three equations with three unknowns, which can be solved to find $\mathbb{E}[T_{1,1}]$, $\mathbb{E}[T_{1,2}]$ and $\mathbb{E}[T_{1,3}]$:

$$E[T_{1,1}] = \alpha \left[\frac{1}{\lambda + \mu} + \frac{\lambda_1}{\lambda + \mu} E[T_{1,1}] + \frac{\lambda_2}{\lambda + \mu} E[T_{1,3}] \right] \quad (19)$$

$$E[T_{1,2}] = \alpha \left[\frac{\lambda_1}{\lambda + \mu} E[T_{1,3}] + \frac{\lambda_2}{\lambda + \mu} E[T_{1,2}] \right] \quad (20)$$

$$E[T_{1,3}] = \alpha \left[\frac{\lambda_1}{\lambda + \mu} E[T_{1,1}] + \frac{\lambda_2}{\lambda + \mu} E[T_{1,2}] \right] \quad (21)$$

Following a similar reasoning,

$$E[T_{2,1}] = \alpha \left[\frac{\lambda_1}{\lambda + \mu} E[T_{2,1}] + \frac{\lambda_2}{\lambda + \mu} E[T_{2,3}] \right] \quad (22)$$

$$E[T_{2,2}] = \alpha \left[\frac{1}{\lambda + \mu} + \frac{\lambda_1}{\lambda + \mu} E[T_{2,3}] + \frac{\lambda_2}{\lambda + \mu} E[T_{2,2}] \right] \quad (23)$$

$$E[T_{2,3}] = \alpha \left[\frac{\lambda_1}{\lambda + \mu} E[T_{2,1}] + \frac{\lambda_2}{\lambda + \mu} E[T_{2,2}] \right] \quad (24)$$

and

$$E[T_{3,1}] = \alpha \left[\frac{\lambda_1}{\lambda + \mu} E[T_{3,1}] + \frac{\lambda_2}{\lambda + \mu} E[T_{3,3}] \right] \quad (25)$$

$$E[T_{3,2}] = \alpha \left[\frac{\lambda_1}{\lambda + \mu} E[T_{3,3}] + \frac{\lambda_2}{\lambda + \mu} E[T_{3,2}] \right] \quad (26)$$

$$E[T_{3,3}] = \alpha \left[\frac{1}{\lambda + \mu} + \frac{\lambda_1}{\lambda + \mu} E[T_{3,1}] + \frac{\lambda_2}{\lambda + \mu} E[T_{3,2}] \right] \quad (27)$$

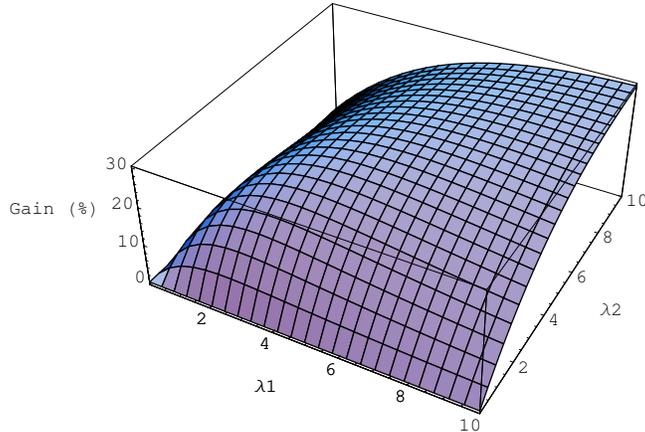


Figure 8: Goodput gain (%), function of λ_1 and λ_2 when $d = 1$.

Once the different $\mathbb{E}[T_{i,j}]$ are found, they are used in Eq. (13) to find the state probabilities $\mathbb{P}(L = 2, N = i)$. Replacing the result of Eqs. (12), (13), and (14) along with the different conditional success probabilities (Eqs. (9)-(11)) into Eq. (8) yields the type-1 goodput of the system, which is a function of λ_1 , λ_2 , μ and d .

6.2. Numerical Results

In this section, we study numerically the goodput gain of network coding relative to the no-coding base case. We are particularly interested in finding how much the goodput of real-time packets flows can be increased by applying network coding. We define γ_{base} as the total goodput of the router without coding, and γ_{coding} as the total goodput of the router with coding. An important remark to be made here is that γ_{base} can be found using Eq. (8), by considering the existence of a single arriving flow instead of two, with parameter $\lambda = \lambda_1 + \lambda_2$. In other words,

$$\gamma_{base} \{ \lambda_1, \lambda_2, \mu, d \} = \gamma_1 \{ \lambda_1 + \lambda_2, 0, \mu, d \}. \quad (28)$$

We define the goodput gain as the relative increase in goodput, i.e., as $(\gamma_{coding} - \gamma_{base}) / \gamma_{base}$.

In the sequel, we fix $\mu = 1$. Fig. 8 provides the goodput gain as a function of the arrival rates λ_1 and λ_2 , for d equal to 1. A major conclusion to be drawn here is that the gain provided by network coding is always maximized

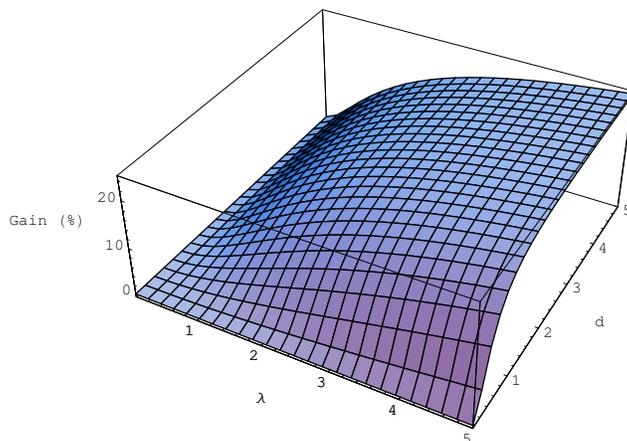


Figure 9: Goodput gain (%) for values of λ between 0 and 5 and values of d between 0 and 5.

when λ_1 is equal to λ_2 . This result is logical, since equal arrival rates provide the most opportunities for coding.

Fig. 9 provides the maximal goodput gain for d between 0 and 5, and λ between 0 and 5. Similarly, Fig. 10 provides the maximal goodput gain for the same range of deadlines and larger values of λ , between 5 and 15.

For a fixed arrival rate λ , the gain increases for an increasing deadline value. A larger deadline provides a higher probability of success for coded packets, hence a positive return on applying network coding. Similarly, for a fixed deadline d , the gain increases for an increasing arrival rate; This can be explained by the more efficient buffer usage that network coding provides; In the base case, every arrival overwrites the buffer. The higher the arrival rate, the more frequent such overwritings occur. On the other hand, with network coding, an arrival of type-1 when the buffer is occupied by a type-2 packet, or vice-versa, does not result in an overwriting, but in a coding operation which maintains both packets. As shown in Fig. 9, the gain provided by network coding remains limited in the operating region ($\lambda < 2$, $d < 1$), where low arrival rates limit the opportunities of coding, and the short deadlines causes most of the coded packets to miss their timeliness requirement anyway. Finally, as conveyed in Fig. 10, the gain of network coding for high arrival rates and large deadlines can reach up to 30%, a substantial improvement compared to the base case scenario. This result is important for highlighting

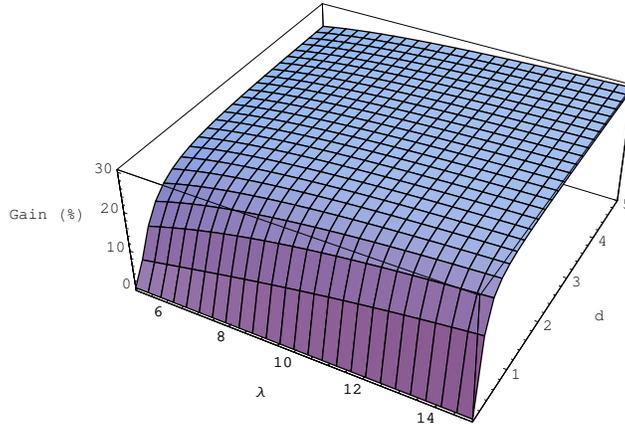


Figure 10: Goodput gain (%) for values of λ between 5 and 15.

the applicability of the proposed approach in more general networks. When the same network coding algorithm was employed for the transmission of real-time packets in a wireless multihop sensor network, the simulation results also showed a maximum performance of 30% for higher values of the traffic load and higher deadlines [2].

A reasoning along the lines of buffer usage efficiency also explains the difference in performance between packet skipping and network coding when the deadline value is increased. As already shown in Fig. 5 and Fig. 6, the gain of packet skipping is limited to the smallest range of deadlines ($d < 1$). In contrast, network coding provides increasing gain for an increasing deadline value, up to reaching a saturation value of 30%. For higher deadlines, lead-time thresholding basically allows any buffered packet to be serviced, thus providing no improvement in buffer usage. On the other hand, the efficient usage of the buffer is maintained in the case of network coding, since coding operations are deadline-oblivious.

7. Joint Coding-Skipping Model

7.1. Goodput Analysis

The following section studies the performance benefits obtained when network coding and packet skipping are concurrently applied. With packet skipping enabled in a network coding setting, instead of removing a waiting packet only once its absolute deadline expires, lead-time thresholding with

threshold θ is applied; if the waiting packet is a native one (a type-1 or type-2 packet), its lead-time is used in the thresholding procedure. If the waiting packet is a coded packet, the largest among the lead-times of the two underlying native packets is used.

The goodput analysis for the joint coding-skipping approach is essentially identical to the analysis of the network coding goodput in Section 6. A number of equations are affected by the introduction of θ ; $E[T]$ takes exactly the expression provided in Eq. (7), in comparison to the simple network coding case where θ was set to 0. The threshold θ is also introduced in the equations of $\mathbb{P}(\text{success}|L = 1)$ and $\mathbb{P}(\text{success}|L = 2, N = 2)$, resulting in:

$$\begin{aligned} \mathbb{P}(\text{success}|L = 1) &= \int_0^{d-\theta} \mu e^{-\mu t} e^{-\lambda_1 t} (1 + \lambda_2 t) e^{-\lambda_2 t} \\ &\times (1 - e^{-\mu(d-t)}) dt. \end{aligned} \quad (29)$$

$$\begin{aligned} \mathbb{P}(\text{success}|L = 2, N = 2) &= \int_0^{d-\theta} \mu e^{-\mu t} e^{-\lambda_1 t} e^{-\lambda_2 t} \\ &\times (1 - e^{-\mu(d-t)}) dt. \end{aligned} \quad (30)$$

Note that the equations giving $\mathbb{P}(\text{success}|L = 2, N = 1)$ and $\mathbb{P}(\text{success}|L = 2, N = 3)$ remain identical to Eq. (29).

Finally, the threshold θ should be taken into account in all $E[T_{i,j}]$ expressions (Eqs. (19)-(27)), by setting α to $1 - e^{-(\lambda+\mu)(d-\theta)}$ instead of $1 - e^{-(\lambda+\mu)d}$.

7.2. Numerical Results

Fig. 11 and Fig. 12 plot the *additional* gain of combining service skipping with network coding, relative to the gains enjoyed from network coding alone. The impact of packet skipping in a joint coding-skipping approach remains limited to the smallest range of deadlines, in accordance with the results for which only packet skipping was employed. However, we observe that the range where the performance of the joint system is noticeable, is smaller than the one of the packet skipping case. A snapshot of this difference can also be seen in Fig. 13. This behavior is consistent with our simulation results in [2] where we observed that the additional gain from packet skipping was

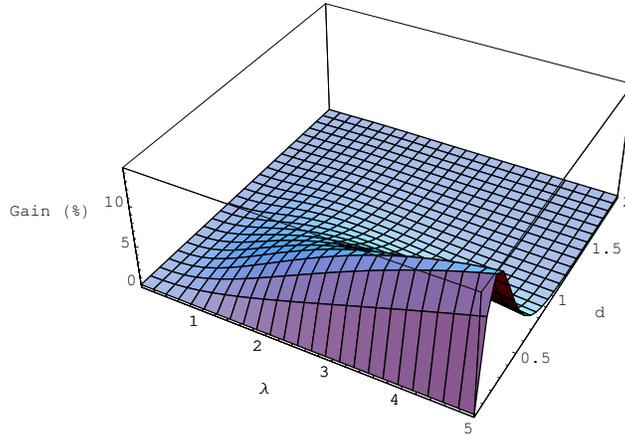


Figure 11: Additional increase in goodput gain (%), for λ between 0 and 5, achieved when adding skipping to network coding.

reaching at most 10-12% and for the lower regime of deadline values. This behavior puzzled us both in our analysis here and also in our simulations. But careful analysis reveals that when network coding is employed, our system naturally removes opportunities for skipping since many native packets, once coded, are transmitted even if they would have been normally dropped with thresholding. This situation is even more true for higher arrival rates since more packets are coded and they are replaced from the buffer faster. Therefore, the opportunities that occur for efficient use of packet skipping become less as the arrival rate is increased.

8. Conclusions

In this paper, we developed an analytical model that captures the impact of packet skipping and network coding on the delivery of real-time packets. First, we developed a model that considers only packet skipping for packets that have associated deadlines. We showed through the analytical model, that an optimal thresholding policy, when applied to the incoming packets at the router, can lead to goodput gains of up to 30%. Subsequently we developed a model for the case where the router employs algebraic coding on the incoming packets. We showed that depending on the operating region,

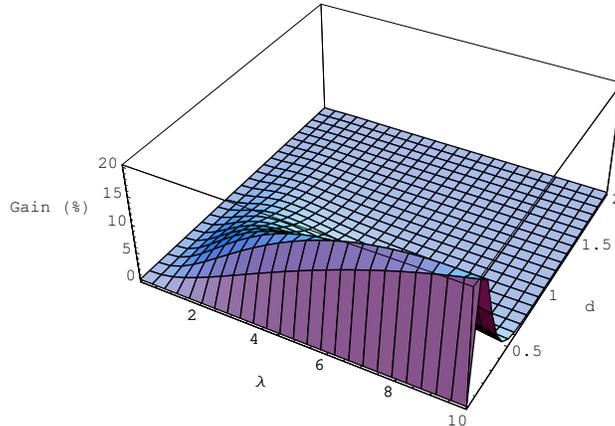


Figure 12: Additional Increase in Goodput gain (%), for λ up to 10, achieved when adding skipping to network coding.

the gain of network coding can reach up to 30%. The ultimate objective of the paper was a router queueing system model that considers jointly network coding and packet skipping in the presence of per-packet deadlines. We provided exact expressions for the stationary goodput of the system and numerical results of the achievable goodput gain. Employing packet skipping in addition to network coding adds up to 15% gain over the collected gains obtained from employing network coding exclusively.

In addition to the novel approach that looks at network coding and packet skipping from a real-time perspective, our work intends to be a stepping stone for future research. Among the topics we foresee as interesting are the qualitative and quantitative studies of network coding and packet skipping in more complex network topologies. Furthermore, we will seek ways to translate some of the results and conclusions of this work to heuristic optimization algorithms for real-life wireless networks.

References

- [1] R. Ahlswede, N. Cai, S.-Y. R. Li, R. W. Yeung, Network information flow, *IEEE Transactions on Information Theory* 46 (4) (2000) 1204–1216.

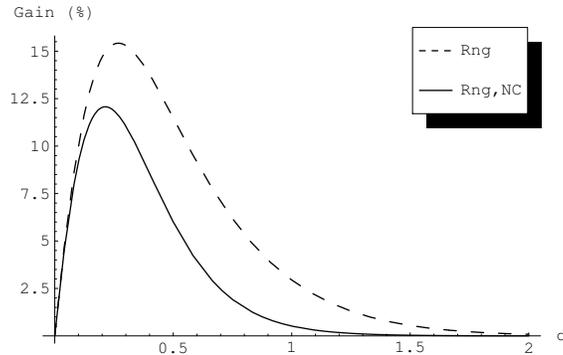


Figure 13: Comparison, for $\lambda = 4$, between gain of skipping and the additional gain it provides when enabled concurrently with network coding.

- [2] M. Aoun, A. Argyriou, P. van der Stok, Performance evaluation of network coding service reneging in iee 802.15.4-based wireless sensor networks, in: European conference on Wireless Sensor Networks (EWSN), 2011.
- [3] L. Kruk, J. Lehoczky, S. Shreve, S.-N. Yeung, Earliest-deadline-first service in heavy-traffic acyclic networks, *The Annals of Applied Probability* 14 (3) (2004) 1306–1352.
- [4] D. Barrer, Queueing with impatient customers and indifferent clerks, *Operations Research* 5 (5) (1957) 644–649.
- [5] D. Barrer, Queueing with impatient customers and ordered service, *Operations Research* 5 (5) (1957) 650–656.
- [6] A. Brandt, M. Brandt, On the $m(n)/m(n)/s$ queue with impatient calls, *Performance Evaluation* 35 (1-2) (1999) 1–18.
- [7] A. Brandt, M. Brandt, Asymptotic results and a Markovian approximation for the $m(n)/m(n)/s+gi$ system, *Queueing Systems* 41 (1-2) (2002) 73–94.
- [8] A. Movaghar, On queueing with customer impatience until the beginning of service, *Queueing Systems* 29 (4) (1998) 337–350.

- [9] A. Movaghar, On queueing with customer impatience until the end of service, *Stochastic Models* 22 (1) (2006) 149–173.
- [10] M. Kargahi, A. Movaghar, Non-preemptive earliest-deadline-first scheduling policy: A performance study, in: *IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, Georgia, ATL, USA, 2005, p. 201210.
- [11] M. Kargahi, A. Movaghar, A method for performance analysis of earliest-deadline-first scheduling policy, *Journal of Supercomputing* 37 (2) (2006) 197–222.
- [12] S. Panwar, D. Towsley, J. Wolf, Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service, *Journal of the ACM* 35 (4) (1988) 832–844.
- [13] J. Lehoczky, Real-time queueing theory, in: *IEEE Real-Time Systems Symposium*, Washington, DC, USA, 1996, p. 186195.
- [14] J. Lehoczky, Using real-time queueing theory to control lateness in real-time systems, *Performance Evaluation Review* 25 (1) (1997) 158–168.
- [15] J. Lehoczky, Real-time queueing network theory, in: *IEEE Real-time Systems Symposium*, San Francisco, CA, USA, 1997, p. 5867.
- [16] B. Doytchinov, J. Lehoczky, S. Shreve, Real-time queues in heavy traffic with earliest-deadline-first queue discipline, *The Annals of Applied Probability* 11 (2) (2001) 332–378.
- [17] P. Parag, J.-F. Chamberland, Queuing analysis of a butterfly network, in: *IEEE International Symposium on Information Theory (ISIT)*, Toronto, Canada, 2008, pp. 672–676.
- [18] D. Shah, M. Médard, J. K. Sundararajan, On queueing in coded networks – queue size follows degrees of freedom, in: *IEEE Information Theory Workshop on Wireless Networks*, Solstrand, Norway, 2007, pp. 1–6.
- [19] A. Eryilmaz, A. Ozdaglar, M. Médard, On delay performance gains from network coding, in: *Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, 2006, pp. 864–870.

- [20] J. Barros, R. A. Costa, J. Widmer, Effective delay control for online network coding, in: IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), Rio de Janeiro, Brazil, 2009, pp. 208–216.
- [21] K. Wu, W. Jia, Y. Yuan, Y. Jiang, Performance modeling of stochastic networks with network coding, in: Workshop on Network Coding, Theory, and Applications (NetCod), Lausanne, Switzerland, 2009, pp. 6–11.
- [22] J. Goseling, R. J. Boucherie, J.-K. Ommeren, Energy consumption in coded queues for wireless information exchange, in: Workshop on Network Coding, Theory, and Applications (NetCod), Lausanne, Switzerland, 2009, pp. 30–35.
- [23] R. Wolff, Poisson arrivals see time averages, *Operations Research* 30 (2) (1982) 223–231.
- [24] R. Wolff, *Introduction to Queueing Theory*, Prentice-Hall, Inc., 1989.
- [25] L. Kleinrock, *Queueing Systems. Volume 1: Theory*, John Wiley & sons, 1975.