Finite-blocklength analysis of the ARQ-protocol throughput over the Gaussian collision channel

Finite-blocklength analysis of the ARQ-protocol throughput over the Gaussian collision channel

Rahul Devassy[1], Giuseppe Durisi[1], Petar Popovski[2], Erik G. Ström[1]

[1]Chalmers University of Technology, 41296 Gothenburg, Sweden
[2]Aalborg University, 9220 Aalborg, Denmark

***Abstract*—We present a finite-blocklength analysis of the throughput and the average delay achievable in a wireless system where i) several uncoordinated users transmit short coded packets, ii) interference is treated as noise, and iii) 1-bit feedback from the intended receivers enables the use of a simple automatic repeat request (ARQ) protocol. Our analysis exploits the recent results on the characterization of the maximum coding rate at finite blocklength and finite block-error probability by Polyanskiy, Poor, and Verdú (2010), and by Yang *et al.* (2013). For a given number of information bits, we determine the coded-packet size that maximize the per-user throughput and minimize the average delay. Our numerical results indicate that, when optimal codes are used, very short coded packets (of length between 50 to 100 channel uses) yield significantly lower average delay at an almost negligible throughput loss, compared to longer coded packets.**

I. Introduction

Next generation wireless communication systems are expected to support real-time data transfer with guaranteed end-to-end delay of at most few milliseconds. This will enable the introduction of services such as [1]

- vehicle to vehicle and vehicle to infrastructure communications for traffic efficiency and safety;
- real-time video processing for augmented reality;
- monitoring of materials, e.g., of buildings to identify potential damages, and monitoring of the environment, for example for agricultural purposes;
- wireless control of industrial plants and smart electricity grids.

A common feature of the services listed above is that they often require the transmission of short packets (no more than hundreds of bits), which need to be correctly decoded at the intended receiver within stringent latency requirements.

Designing wireless communication systems able to support such services is challenging because most of the results available within the field of wireless communication theory are asymptotic in the packet length. Indeed, the classic performance metric used in wireless communication theory is the *Shannon capacity* [2], which is the largest data rate at which reliable communication (i.e., communication with arbitrarily low error probability) is possible. This metric is asymptotic in the following sense: for a given rate below the Shannon capacity, arbitrarily low error probability can be achieved only using sufficiently long (coded) packets, i.e., introducing sufficiently long delays.

In spite of its asymptotic nature, Shannon capacity (and its extension to nonergodic channels—the outage capacity [3, p. 2631], [4]) has been proven useful to design the current wireless systems. The reason is that the delay constraints in current systems are typically above 10 ms, which, for the frequency-bandwidth values presently used, allows for long packets.

However, when the transmitted packets are short, channel capacity is a poor benchmark. In this scenario, the fundamental performance limit is instead the *maximum achievable rate* $R^*(n,\epsilon)$ at a given packet length n and packet error probability ϵ. This quantity, which is proportional to the largest amount of information bits k that can be mapped into a packet of n coded bits, under the constraint that the information bits are recovered at the receiver with probability no smaller than $1-\epsilon$, has been recently characterized in [5]–[7] for both AWGN and fading channels.

The objective of this paper is to leverage the results obtained in [5]–[7] to investigate the following question: *how should one design throughput optimal multiple-access schemes in the presence of latency constraints?* This question has been posed before, often in the context of cross-layer optimization (see e.g, [8]). However, most of the available results are for specific combinations of modulation and coding schemes [9] or rely on asymptotic information-theoretic performance metrics [10].

In this preliminary investigation, we shall focus on a simple system model, namely, a wireless communication system where several *uncoordinated* users transmit short coded packets using frequency hopping and a simple automatic repeat request (ARQ) protocol. This setup, which is closely related to the one known in the literature as *slotted Gaussian collision channel with feedback* [11], [12], is particularly relevant for machine-type communication systems involving a very large number of devices. We consider both the case when the channel among each users is impaired by additive Gaussian noise only, and the quasi-static fading case, where the fading gains are random but stay constant over the duration of each packet.

An analysis of the throughput achievable over this channel has been previously conducted in [12] for the case of asymptotically large packet length, by relying on the capacity-versus-outage formulation. Our analysis is instead non-asymptotic. Specifically, we use the results on the characterization of $R^*(n,\epsilon)$ reported

This work was partly funded by the Swedish Research Council under grant 2012-4571. The simulations were performed in part on resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE.

in [5]–[7] to analyze the throughput achievable over this channel. For a given number of information bits, we determine the coded packet size that maximizes the throughput and minimizes the average transmission delay. We also investigate how the maximum throughput and the minimum average delay behave as a function of the number of information bits. Our numerical results indicate that, when optimal codes are used, very short coded packets (of length between 50 to 100 channel uses) corresponding to about 100 information bits yield significantly lower average delay at an almost negligible throughput loss compared to longer packets containing more information bits (1000).

II. System Model

We analyze a slotted Gaussian collision channel with feedback [12], over which N_u transmitter-receiver pairs operate concurrently. For the fading scenario, we consider communications over a time-frequency selective fading channel with coherence time T_c and coherence bandwidth W_c. The available bandwidth $W > W_c$ is divided into $n_f = W/W_c$ non-interferring frequency bands. For simplicity, we shall assume in the following that n_f is an integer. For each slot, each user chooses a frequency band uniformly at random and independently from the other users, and transmits over this band a coded packet consisting of n complex symbols (corresponding to n channel uses) of duration $n/W_c < T_c$ seconds. These assumptions guarantee that the fading channel stays constant over the duration of each coded packet.

The received vector $\mathbf{y} \in \mathbb{C}^n$ corresponding to the coded packet $\mathbf{x}_1 \in \mathbb{C}^n$ transmitted by user 1 during one (arbitrary) packet transmission slot is given by

$$\mathbf{y} = h_1\mathbf{x}_1 + \sum_s h_s\mathbf{x}_s + \mathbf{w}. \tag{1}$$

Here, h_s denotes the fading coefficient corresponding to user s, the index s spans the set of interfering users (i.e., of users that chose the same frequency band as user 1 for transmission), and $\mathbf{w}$ models the additive noise vector, whose entries are independent and identically distributed circularly symmetric complex Gaussian random variables with unit variance.

At the intended receiver, which is assumed perfectly aware the frequency band chosen by the corresponding transmitter,[1] but which ignores the choice of the other (unintended) users, decoding is attempted. A binary feedback about the status of the decoding operation is sent back to the transmitter. If the feedback indicates a decoding failure, the transmitter repeats the same coded packet over the next packet transmission slot, after having selected a different frequency band. This allows for reliable data transmission in spite of interference and deep-fade events. If the feedback indicates decoding success, the next coded packet is transmitted.

Each coded packet corresponds to k information bits (we assume that all users need to deliver similar payloads). Furthermore, each user maps the information bits into coded bits independently from the other users, i.e., no coordination among users is assumed.

To simplify the analysis, we shall also assume what follows: i) Each user has an infinite number of information bits to transmit (*full-buffer* assumption) so that as soon as the transmission of the current packet is stopped because decoding is successful, the transmission of the next packet is started. ii) The feedback is instantaneous and error free. iii) Interference resulting from several users contending the same medium is treated as additive Gaussian noise. iv) All users transmit at the same power ρ. v) The fading coefficients $\{h_s\}$ are independent and identically distributed and perfectly known to the receiver.

Assumptions iii) and iv) imply that, given the fading coefficients $\{h_s\}$, the second and third term on the right-hand-side of (1) can be jointly modeled as a circularly symmetric Gaussian random variable with variance $1 + \rho\sum_s |h_s|^2$. Furthermore, under the assumptions listed above, the system is symmetric with respect to any user. For simplicity, we shall then take user 1 as the reference user in the remainder of the paper.

It is appropriate at this point to comment on the differences between the system model analyzed in [12] and the one in this paper. In [12] it is assumed that all users transmit at a given rate R. As the packet length is assumed large, the probability of erroneous packet detection ϵ is given by the outage probability[2]

$$\epsilon = \Pr\{\log_2(1+\rho_{\text{int}}) < R\} \tag{2}$$

where ρ_{int} is the signal-to-interference-and-noise ratio

$$\rho_{\text{int}} = \frac{\rho\,|h_1|^2}{1+\rho\sum_s |h_s|^2}. \tag{3}$$

In our setup, we fix instead the number of information bits k that each user needs to transmit. The packet length n becomes an optimization parameter that can be set so as to maximize the throughput. As the resulting packet size is (possibly) small, (2) is not necessarily valid.

To make packet repetition effective, we are interested in the scenario where different packets from the same user experience different fading realizations. As the packet length is an optimization parameter, we cannot a priori guarantee that it will be as large as the channel coherence time. Furthermore, as we are interested in minimizing the average delay, interleaving or inserting guard intervals is not an appealing solution. Instead, we achieve diversity through frequency hopping. In contrast, the scheme proposed in [12] considers time-hopping only, according to the slotted-Aloha paradigm. Furthermore, the role of time-hopping in [12] is to mitigate collisions, not to provide diversity.

The non-fading (AWGN) scenario is readily obtained from (1) by assuming that the channel gains in (1) are deterministic.

III. Maximum Coding Rate at Finite Blocklength

In this section, we briefly review the recent results on the characterization of the maximum coding rate $R^*(n,\epsilon)$ at finite

[1] This is achieved in practice by letting each user transmit according to a predetermined frequency-hopping pattern.

[2] Note that (2) holds independently on whether the interference is assumed Gaussian or not. Indeed, a standard typicality argument shows that a receiver that treats interference as noise achieves (2).

blocklength and finite error probability [5]–[7] that we shall need for our analysis in Section IV.

A. AWGN channels

To define $R^*(n,\epsilon)$, we shall focus on the single-user AWGN case, for which the input-output is given by

$$\mathbf{y} = \mathbf{x} + \mathbf{w}. \tag{4}$$

An (n, M, ϵ) code for the AWGN channel (4) consists of:

1) an encoder f: $\{1,\dots,M\} \mapsto \mathbb{C}^n$ that maps the message $J \in \{1,\dots,M\}$ into a codeword $\mathbf{x} \in \{\mathbf{c}_1,\dots,\mathbf{c}_M\}$ satisfying power constraint

$$\|\mathbf{c}_j\|^2 \le n\rho, \quad j = 1,\dots,M. \tag{5}$$

2) A decoder $g : \mathbb{C}^n \mapsto \{1,\dots,M\}$ that satisfies the average error probability constraint (packet error rate constraint)

$$\Pr[g(\mathbf{y}) \neq J] \le \epsilon \tag{6}$$

where $\mathbf{y}$ is the channel output induced by the transmitted codeword according to (4), and J is uniformly distributed over the message set $\{1,\dots,M\}$.

The maximum achievable rate is defined as

$$R^*(n,\epsilon) = \sup\left\{\frac{\log_2 M}{n} \;:\; \exists (n, M, \epsilon) \text{ code}\right\}. \tag{7}$$

Nonasymptotic upper and lower bounds on $R^*(n,\epsilon)$ are reported in [5, Eq. (218)]. The bounds are remarkably tight. Furthermore, their analysis for $n \to \infty$ allows one to establish the following asymptotic characterization for $R^*(n,\epsilon)$ [5, Eq. (296)], [13][3]

$$R^*(n,\epsilon) = C(\rho) - \sqrt{\frac{V(\rho)}{n}}Q^{-1}(\epsilon) + \frac{\log_2 n}{2n} + \mathcal{O}(1). \tag{8}$$

Here,

$$C(\rho) = \log_2(1+\rho) \tag{9}$$

is the channel capacity,

$$V(\rho) = \left(1 - \frac{1}{(1+\rho)^2}\right)\log_2^2 e \tag{10}$$

is the so-called channel dispersion, and $Q^{-1}(\cdot)$ stands for the inverse of the Gaussian Q-function. Finally, the notation $f(n) = \mathcal{O}(g(n))$, $n \to \infty$ means that $\limsup_{n\to\infty} |f(n)/g(n)| < \infty$. The asymptotic expansion (8), which is shown in [5] to be accurate already at packet sizes as small as 50, suggests the following approximation for the minimum packet error rate incurred in the transmission of $k = \log_2 M$ information bits over the AWGN channel (8) using coded packet spanning n channel uses:

$$\epsilon(n,k) \approx Q\left(\frac{nC(\rho) + 0.5\log_2 n - k}{\sqrt{nV(\rho)}}\right). \tag{11}$$

B. Quasi-static fading channels

For the quasi-static case, the input-output relation is given by

$$\mathbf{y} = h\mathbf{x} + \mathbf{w}. \tag{12}$$

We shall assume that the receiver is perfectly aware of the realizations of the fading channel h. Under this assumption, the decoder g depends on both the channel output and the fading realizations. In practice, knowledge of h is acquired by transmitting extra pilot symbols. The resulting overhead (neglected in the present paper) has been partially characterized in [14]. Bounds on $R^*(n,\epsilon)$ for the quasi-static case were recently reported in [6], [7]. These bounds yield the following asymptotic approximation for $R^*(n,\epsilon)$

$$R^*(n,\epsilon) = C_\epsilon + \mathcal{O}\left(\frac{\log_2 n}{n}\right). \tag{13}$$

Here, C_ϵ denotes the outage capacity, i.e.,

$$C_\epsilon = \sup\{R : \Pr[\log_2(1+\rho\,|h|^2) < R] < \epsilon\}. \tag{14}$$

From (8) and (13), we conclude that $R^*(n,\epsilon)$ approaches the asymptotic ($n \to \infty$) limit much faster in the quasi-static case compared to the AWGN case. Since the quasi-static fading channel is conditionally Gaussian given the channel gain h, the following approximation on the minimum packet error rate incurred in the transmission of $k = \log_2 M$ information bits using a length n coded packet turns out to be accurate:

$$\epsilon(n,k) \approx \mathbb{E}\left[Q\left(\frac{nC(\rho\,|h|^2) + 0.5\log_2 n - k}{\sqrt{nV(\rho\,|h|^2)}}\right)\right]. \tag{15}$$

Here, the expectation is over the channel gain $|h|^2$. This approximation is a minor improvement over the ones previously reported in [15], [7], [16].

IV. Finite Blocklength Analysis

Our throughput analysis follows closely [12]. Using the renewal-reward theorem [17] and (15) we conclude that the overall system throughput η, measured in bits per second per Hertz (or bits per channel use), corresponding to the transmission of coded packets of length n is given by

$$\eta = N_u \frac{k}{n}\left[1 - \epsilon_{\text{int}}(n,k)\right] \tag{16}$$

where, because of (15), the packet error rate in the presence of interference $\epsilon_{\text{int}}(k,n)$ can be well-approximated by

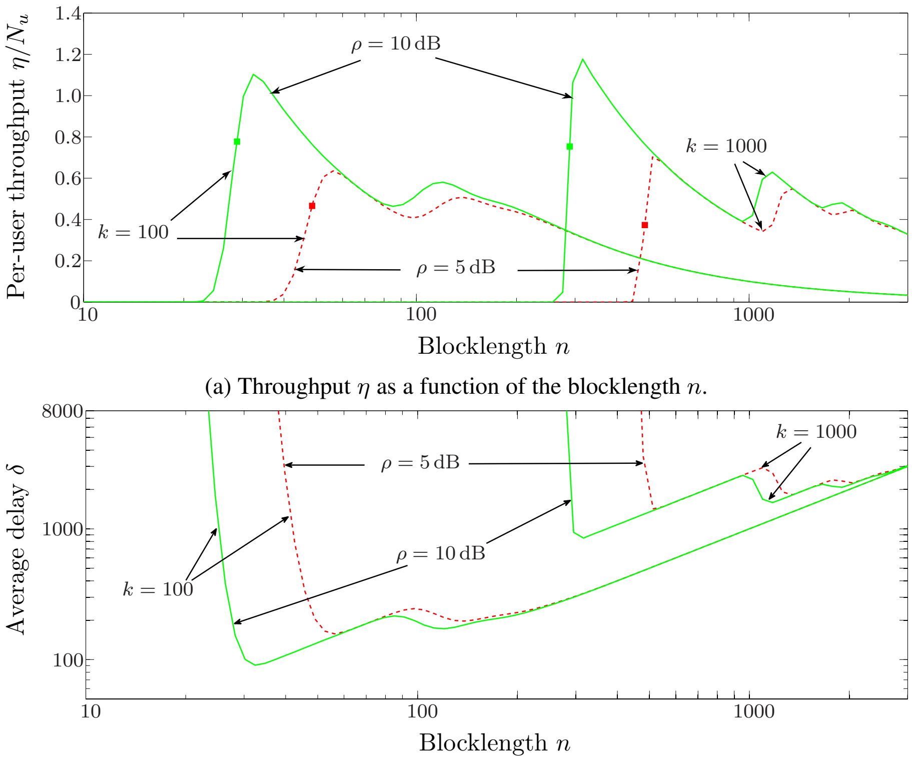
$$\epsilon_{\text{int}}(n,k) \approx \mathbb{E}\left[Q\left(\frac{nC(\rho_{\text{int}}) + 0.5\log_2 n - k}{\sqrt{nV(\rho_{\text{int}})}}\right)\right]. \tag{17}$$

Here, the averaging is performed with respect to the random variable ρ_{int} defined in (3). The corresponding average delay (measured in number of channel uses) is given by

$$\delta = \frac{n}{1 - \epsilon_{\text{int}}(n,k)}. \tag{18}$$

This expression holds under the assumption of unlimited number of retransmissions. In the next section, we shall numerically

[3] The results reported in [5, Eq. (296)], [13] pertain to the *real* Gaussian channel. The extension to the complex case in (8) follows by identifying a complex Gaussian channel with blocklength n with a real AWGN channel with the same signal-to-noise ratio and blockength $2n$.



(a) Throughput η as a function of the blocklength n.

(b) Average delay δ as a function of the blocklength n.

Fig. 1: AWGN collision channel; $N_u = 50$; $n_f = 50$.

optimize both throughput and average delay as a function of the blocklength n for various values of information bits k and transmitted power ρ.

V. Numerical Results

Throughout, we shall assume $N_u = 50$ users and $n_f = 50$. This corresponds to an average load per frequency band of 1 user. We consider two SNR values, namely $\rho = 5\,\mathrm{dB}$ and $\rho = 10\,\mathrm{dB}$, two values for the number of information bits per packet $k = 100$ and $k = 1000$, and both the AWGN and the quasi-static case.

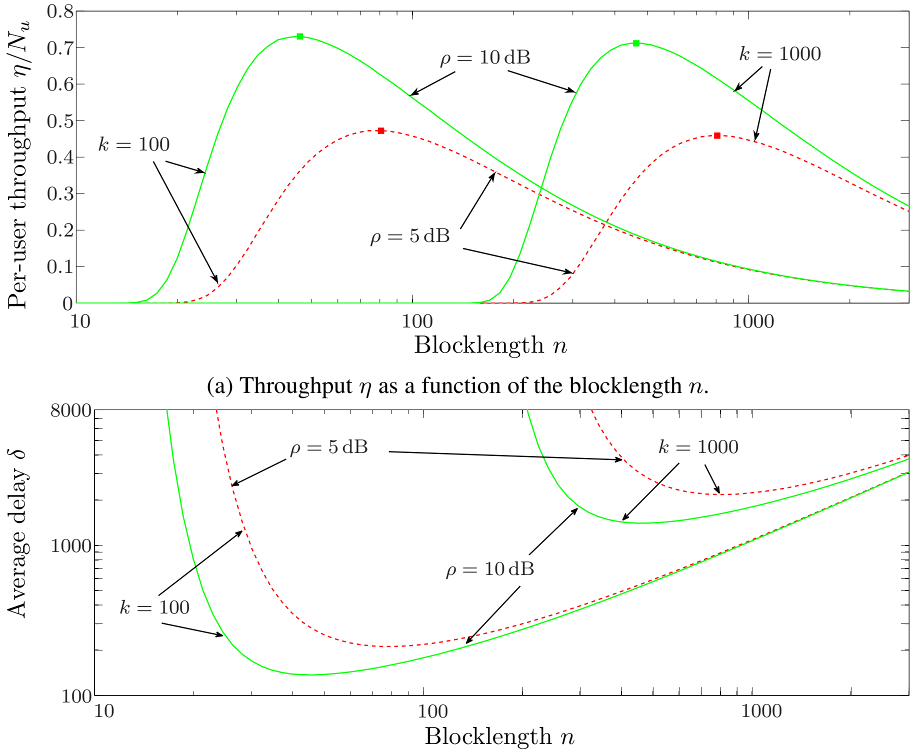
1) AWGN: Fig. 1a shows the behavior of the per-user throughput η/N_u as a function of the blocklength n. Let us focus for simplicity on the case $k = 100$ and $\rho = 5\,\mathrm{dB}$ (leftmost dashed red curve). The throughput maximizing blocklength value is about 50 channel uses. This value also minimizes the average delay (see Fig. 1b). Choosing n so that the ratio k/n equals the rate that maximizes the throughput (with interference treated as noise) in the asymptotic regime of long packets (see [12]), results in the throughput value indicated by the square marker, i.e, a throughput loss of about 26%. If we increase the number of information bits from 100 to 1000, we observe a slight increase in the throughput (about 10%), but a much more significant increase in the average delay (about 9 times larger). As expected, a high value of the transmit power ρ results in a larger throughput and a smaller average delay, because the corresponding optimal packet size is smaller.

2) Quasi-static: The channel gains $\{h_s\}$ are modeled as independent circularly symmetric complex Gaussian random variables with unit variance (Rayleigh fading). We see from Fig. 2a that the markers corresponding to the throughput achievable by choosing n so that the ratio k/n approach the optimal rate for the asymptotic regime of long packets are close to the actual maximum. This confirms the observation reported in [7] that the outage capacity is a sharp proxy for the finite-blocklength fundamental limits of quasi-static channels. Differently from the Gaussian channel, increasing the number of information bits from 100 to 1000 results in a throughput reduction (about 3% for $\rho = 5\,\mathrm{dB}$). This is because the approximation on $R^*(n, \epsilon)$ resulting from (15) converges to the outage capacity from above for the power levels considered in the plot. As shown in Fig. 2b, this modest throughput degradation comes with a much more significant increase in the average delay (about 10 times larger for $\rho = 5\,\mathrm{dB}$)

VI. Conclusions and Future Directions

We analyzed the throughput and the average delay achievable over a slotted Gaussian collision channel with feedback, as a function of the number of information bits per coded packet and the coded-packet size. Our results show that at moderate SNR values, it is preferable to use short coded packet carrying few information bits. Indeed, the small throughput increase achievable (for the AWGN case) using longer coded packets comes with a significant increase in the average delay.

Our analysis was focussed on a simple ARQ protocol. To generalize our results to more sophisticated schemes such as



(a) Throughput η as a function of the blocklength n.

(b) Average delay δ as a function of the blocklength n.

Fig. 2: Quasi-static collision channel; $N_u = 50$; $n_f = 50$.

repetition-packet diversity and more general forms of hybrid-ARQ, one needs to find suitable finite-blocklength approximations to be used instead of (17). For the case when the users transmit to a centralized receiver, treating interference as noise is a clearly suboptimal strategy, and more sophisticated protocols can be used to improve the throughput. One such example is coded slotted Aloha [18], [19] where a packet-oriented code is used to create dependency among multiple packets, and collisions between packets are resolved at the receiver through iterative interference cancellation. A finite blocklength analysis of this protocol may shed light on its suitability for applications with stringent delay constraints.

References

[1] METIS project, Deliverable D1.1, "Scenarios, requirements and KPIs for 5G mobile and wireless system," Tech. Rep., Apr. 2013.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

[3] E. Biglieri, J. G. Proakis, and S. Shamai (Shitz), "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[4] L. H. Ozarow, S. Shamai (Shitz), and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 359–378, May 1994.

[5] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[6] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static SIMO fading channels at finite blocklength," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013, pp. 1531–1535.

[7] ——, "Quasi-static MIMO fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, Nov. 2013, submitted for publication. [Online]. Available: http://arxiv.org/abs/1311.2012

[8] M. van der Schaar and D. Turaga, "Cross-layer packetization and retransmission strategies for delay-sensitive wireless multimedia transmission," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 185–197, Jan. 2007.

[9] A. Tandon, M. Motani, and V. Srivastava, "On the impact of channel coding on average packet delay in a multiuser environment," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Shanghai, China, Apr. 2013, pp. 499–504.

[10] E. Setton, T. Yoo, X. Zhu, A. Goldsmith, and B. Girod, "Cross-layer design of ad hoc networks for real-time video streaming," *Wireless Communications, IEEE*, vol. 12, no. 4, pp. 59–65, Aug. 2005.

[11] G. Caire, E. Leonardi, and E. Viterbo, "Modulation and coding for the Gaussian collision channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 2007–2026, Jun. 2000.

[12] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.

[13] V. Y. F. Tan and M. Tomamichel, "The third-order term in the normal approximation for the awgn channel," Nov. 2013. [Online]. Available: http://arxiv.org/abs/1311.2337

[14] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Diversity versus channel knowledge at finite block-length," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Lausanne, Switzerland, Sep. 2012, pp. 572–576.

[15] ——, "Block-fading channels at finite blocklength," in *Proc. IEEE Int. Symp. Wirel. Comm. Syst. (ISWCS)*, Illmenau, Germany, Aug. 2013.

[16] E. MolavianJazi and J. N. Laneman, "On the second-order coding rate of block fading channels," in *Proc. Allerton Conf. Commun., Contr., Comput.*, Monticello, IL, USA, Oct. 2013, to appear.

[17] R. Wolff, *Stochastic modeling and the theory of queues*. Upper Saddle River, NJ, U.S.A.: Prentice Hall, 1989.

[18] E. Paolini, G. Liva, and M. Chiani, "High throughput random access via codes on graphs: Coded slotted ALOHA," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kyoto, Japan, Jun. 2011.

[19] ——, "Coded slotted ALOHA: A graph-based method for uncoordinated multiple access," Jan. 2014. [Online]. Available: http://arxiv.org/abs/1401.1626