CCIT Report #659 November 2007

On Queueing and Multi-Layer Coding

Avi Steiner, Shlomo Shamai (Shitz)

DRAFT

Abstract

A single-server queue concatenated with a multi-level channel encoder is considered. The main focus of this work is on minimization of the average delay of a packet from entering the queue until completion of successful service. Tight bounds are derived for the average delay for different numbers of coded layers. Numerical optimization is applied to find the optimal resource allocation minimizing the average delay. Delay bounds are also derived for continuous layering (single user broadcast approach). The optimizing power distribution of the minimal delay is approximated, and numerically evaluated. It is demonstrated that code layering may give pronounced performance gains in terms of delay, which are more impressive than those associated with throughput. This makes layering more attractive when communicating under stringent delay constraints.

I. Introduction

In classical information theory, a maximal transmission rate is sought for, under a power constraint, assuming an infinite backlog of information waiting for transmission (Shannon capacity). In network theory the input data is some random process which controls writing to a queue, and the output transmission (service) is another random process. In this setting the design goal of the transmission process concentrates on minimizing the queue delay for the input data, under some power constraint. In presence of stringent delay constraints on input data transmission the design of the data queue and transmission algorithm cannot be separated, as maximal throughput is no longer the issue. This conceptual gap between information theory and network theory can be overcome by jointly solving a common problem of minimizing the delay for some input random process and a power/rate control constraint. This is also known as cross-layer optimization, since it involves joint optimization of two layers of the seven layer OSI model. Further inherent gaps between information theory and network theory are discussed in detail in [1], [2], [3].

The tradeoff between delay and throughput has been considered in several contributions [4], [5], [6], [7] and more. Single server queue throughput analysis may be found in [5], for an additive white Gaussian noise (AWGN) channel with different service time distributions. In [6] the channel model is a fading channel with channel state information (CSI) known at both transmitter and receiver ends. Power allocation for a single user for minimizing average delay is also considered in [8], where the fading channel has a changing signal to interference ratio (SIR), depending on the number of users transmitting simultaneously. A communication scheme that is suitable for mixed delay-constrained and non-delay constrained services simultaneously is suggested in [9]. This is obtained by a transmission scheme based on sub-channel grouping together with different power control policies.

Single-user queueing and channel coding for a block fading channel when CSI is available only at the receiver end is discussed in [10], [11]. In [10], optimal rate and power allocation are derived for a single level encoder at the transmitter, where the maximal throughput achievable is also known as the outage capacity. Power and rate are jointly optimized to minimize overall delay, which is the delay between a packet arriving at the queue and being successfully decoded (including retransmission on outage events).

When a separate control can be applied for every transmission burst, the overall average performance may be improved by dynamically controlling the rate, power, transmission algorithm, etc.. A common dynamic optimization framework is dynamic programming [12]. In [13], transmission over a time varying channel, with delay and peak-power constraints, is optimized via dynamic programming. Power allocation policies, as a function of the queue size and channel state are investigated [13]. In [7], the authors use dynamic programming to compute the optimal power allocation for a single user (single server) fading channel, with CSI at transmitter and receiver. Two transmission models are considered there, the first corresponds to fixed length variable rate codewords, and the second corresponds to variable length codewords. The authors of [14] derive optimal power allocation for a wireless fading channel. That is an optimal policy for every channel state and queue state is presented, numerical computations via dynamic programming demonstrate results. Maximization of data throughput for an energy and time constrained transmitter sending data over a fading channel is considered in [15]. A dynamic programming formulation that leads to an optimal closed-form transmission scheduling is obtained. The result is extended to the problem of minimizing the energy required to send a fixed amount of data over a fading channel given deadline constraints [15]. Optimal power allocation and admission control via dynamic programming in context of satellite communications is also presented in [16]. A dynamic programming formulation for computing optimal power control, source coding, and channel coding policies when the source traffic has tight delay constraints is presented in [17].

A general dynamic programming framework for optimal cross-layer adaptation of single-user wireless channels and a stochastic approximation formulation for distributed power and admission control in ad-hoc networks for time-varying channels is discussed in [18]. Random channel environments are discussed in [19]. Generally stating, the parameters for dynamic optimization depend on the system flexibility and dynamic computation capabilities.

Numerous works consider cross-layer optimization for multiple users (multi-server queue), each holding a queue of data, and encountering collisions or other time varying conditions, [20], [21], [22], [23], [24], [25], [26], [27], [28] and more. For such systems there are many retransmission protocols and coordination algorithms. In this context a multiple-access (MAC) Gaussian channel is analyzed in [21], where an information theoretic view of some basic protocols based on the hybrid-ARQ (Automatic Repeat reQuest) are considered. In [25] an ALOHA system is studied, where multiple users transmit synchronously over a time-slotted multiple-access channels. When a collision occurs the users need to retransmit their data. Capacity of time-slotted ALOHA was studied in [22]. Different scheduling schemes are considered in [23] for reduced delay on the expense of throughput and vise-versa, for a single antenna broadcast-fading channel. The transmission there is assumed to be packet based, and average delay and its variance are derived. In [24], throughput-delay trade-off in energy constrained multi-user random wireless network with uniformly distributed nodes is considered, and the optimal tradeoff between average energy-per-bit and delay scaling is presented there. The scheme in [25] considers a broadcast coding scheme, which allows decoding of partial information in case of a collision, and full-decoding in absence of collision. This approach of broadcast coding for the multiple-access channel was first considered in [29].

In this work, we consider a single server queue followed by a channel encoder, which can perform multi-level coding. The channel model is a block fading channel, where CSI is available at the receiver end only. In this case throughput gains may be obtained by performing finite level coding or continuous layering (single-user broadcast approach) [30], [31], [32] and [33].

Stringent delay constraints are common in many applications such as voice/video transmission. In this channel model the transmission block though still large (as to give rise to the notion of reliable communication [34]) is much shorter than the dynamics of the fading process. This scenario is approximated by assuming that the channel fading coefficients are fixed for every block. The notion of capacity versus outage was introduced and discussed in [34] and [35, see references therein].

The focus of this paper is on the overall delay assuming that the input data is kept in a queue and has a fixed finite rate. Optimal rate and power allocation are derived for a multi-level channel encoder, and delay gains of layering are compared to layering throughput gains.

The single-user broadcasting approach hinges on the broadcast channel, which was first explored by Cover [36], [37]. In a broadcast channel a single transmission is directed to a number of receivers, each enjoying possibly different channel conditions, reflected in their received signal-to-noise ratio (SNR). The Gaussian broadcast channel with a single transmit antenna coincides with the classical physically degraded Gaussian broadcast channel, whose capacity region is well known [37],[38], [39]. Single-user broadcasting may be interpreted as hierarchical coding via multi-level coding (MLC) [40], [41], [42], [43].

In general, layered coding includes different data for each layer. This might suggest that when combining the transmitter with the queueing system that the data for each layer is to be stored in a different queue. This is also the case in the general broadcasting problem without common information, where a queue is allocated to every user. However, for single-user communications, a single common queue for all layers is preferable, since it gives the flexibility of dynamically allocating data to layers before every transmission block.

A joint optimization of queueing and multi-user communications is considered in [28] and [44]. The authors derive optimal adaptive joint power control and rate allocation policies which maximize system delay and throughput for multi-access and broadcast fading channels. In both settings every user has his own queue. In the channel model, CSI is assumed to be known for all users at the transmitter. Moreover, for delay optimization it is assumed that every user captures the channel ergodicity, and that all users have the same fading random process (symmetry assumption [44, Section 4.A]). In this setting every user experiences all possible fading realizations. These results are incompatible with the single-user broadcast approach [30], where every user is associated with a channel fading power. That is, every layer (user) is related to a channel fading amplitude, or range of amplitudes. Hence the channel distribution associated with a layer is only a random phase, which means that the distribution of CSI among users is different.

In the concatenation of a queue and multi-level encoder every block transmission consists of multi layer data, which is decoded partially or completely, depending on the fading conditions. The better the channel conditions, the more layers are decoded. Since the transmitter has no access to CSI, a feedback acknowledge (ACK) channel is required to specify which of the layers were decoded. For each ACK the corresponding data can be deleted from the queue. Layers which were not ACKed remain in the queue and are retransmitted. This is equivalent to batch processing, where a layer is interpreted as a batch job, and each service may include processing of several jobs [45].

The structure of this paper is as follows. In section II the channel model is presented. Then the queue model used for the analysis is peresented in section III. In section IV, simple upper and lower bounds on the average delay are presented. In section V tight upper and lower delay bounds are derived for queueing and two level code layering. The results are extended to K-level code layering in section VI, where tight upper and lower delay bounds are derived, closely approximating the exact average delay value. Section VII further extends the delay bounds to general continuum layering, namely the broadcast approach. The numerical results of finite-level code layering and broadcasting are presented in section VIII. Finally, section IX includes the summary and conclusion.

II. CHANNEL MODEL

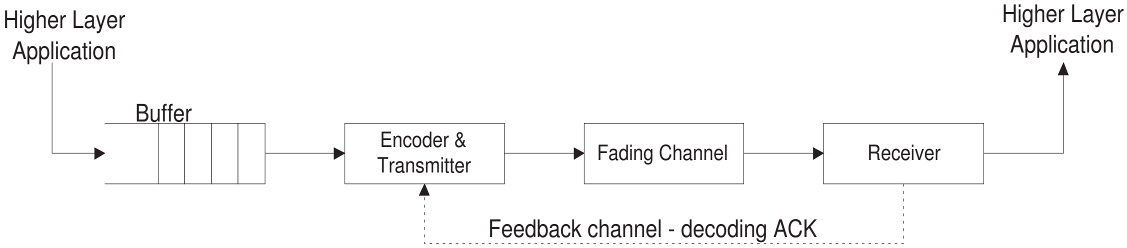

Fig. 1. A block diagram of a communication system including a queue buffer before the transmitter.

Consider the following single-input single-output (SISO) channel,

$$\mathbf{y} = h\mathbf{x} + \mathbf{n} \ , \tag{1}$$

where **y** is a received vector, of length N, which is also the transmission block length, **x** is the transmitted vector. **n** is the additive noise vector, with elements that are complex Gaussian i.i.d with zero mean and unit variance denoted $\mathcal{CN}(0, 1)$, and h is the (scalar) fading coefficient. The fading h is assumed to be perfectly known at the receiver end only. The transmitter has no channel state information (CSI). The power constraint is given by $P = E|x|^2$. E stands for the expectation operator.

Figure 1 illustrates a system including a buffer (queue) for source data, followed by a channel encoder and transmitter. The input data comes from a higher layer application. It enters the queue in a fixed rate λ, and taken out of the queue according to channel encoder scheduler. The queue represents here the networking layer, and the transmitter represents the physical layer. In a single-level coding (outage) approach every transmission block as of the same length and rate. An ACK is returned every successful decoding, and a NACK is conveyed back every time the channel conditions do not allow decoding (outage event). In multi-level coding there is a separate ACK/NACK feedback for every layer. Layers which were NACKed remain in the queue and are scheduled for retransmission.

III. Queue Model - the zero-padding queue

We focus here on a queue model, which allows transmission even when the queue is almost empty and a transmission frame can be created by zero padding the current data in the queue to construct a valid frame for the channel encoder. We introduce the queueing time and waiting time, which is defined as the time spent in a queue from the arrival until taken out of the queue. Queueing time is also the overall delay, defined as the time from arrival to the queue until completion of service (successful transmission).

The waiting time in the queue may be analyzed in the embedding points, at the beginning of every slot. We assume that the input data rate is λ [bits/channel use]. The input arrival epochs is assumed to be deterministic, in-between the embedding points.

The waiting time in a queue may be obtained from the queue size, by normalizing the queue size by the inverse of the input rate λ, as stated by Little's theorem [46]. The queue size is defined by the following equation

$$W_{n+1} = \begin{cases} W_n + N\lambda_{n+1} - NR_n & W_n + N\lambda_{n+1} - NR_n \geq 0 \\ 0 & W_n + N\lambda_{n+1} - NR_n < 0 \end{cases} \tag{2}$$

where N is the block length (number of channel uses between slots), λ_{n+1} is a random variable (RV) of the input rate, which is either a Poisson process at rate λ or a deterministic fixed rate λ, R_n is the transmission rate random variable. Notice that in an outage approach R_n is a fixed R with probability p, and 0 with probability $1-p$. This waiting time equation is also analyzed in [11, ch. 5] for a deterministic arrival process and an outage approach, where bounds on the average waiting time are derived. For convenience, the queue size equation is normalized by the block length N, and we get the queue equation, known also as the Lindley equation [47],

$$w_{n+1} = \begin{cases} w_n + \lambda_{n+1} - R_n & w_n + \lambda_{n+1} - R_n \geq 0 \\ 0 & w_n + \lambda_{n+1} - R_n < 0 \end{cases} \tag{3}$$

where λ_{n+1} is a random variable of the normalized input rate λ, and R_n is the normalized transmission rate random variable. And w_n is now the queue size in units of blocks of data corresponding to N arrivals to the queue. In an outage approach,

$$R_n = \begin{cases} R & w.p. \quad p \\ 0 & w.p. \quad 1-p \end{cases} . \tag{4}$$

From here on, the queue equations will be normalized following (3). For completeness of definition, we state the queueing time equation, which is the overall system delay (overall time spent in the system), for the zero-padding queue. When interested in the overall delay, one has to consider the additional delay of service time, beyond the waiting time in the queue. This is formalized in the next normalized queueing time equation,

$$q_{n+1} = \begin{cases} q_n + \frac{\lambda_{n+1}}{\lambda} - \frac{R_n}{\lambda} & q_n - \frac{R_n}{\lambda} \geq 0 \\ \frac{\lambda_{n+1}}{\lambda} & q_n - \frac{R_n}{\lambda} < 0 \end{cases} \tag{5}$$

where q_n is the momentary queueing time at slot n, and λ_{n+1}, R_n are defined below (3).

In the outage approach it might be desirable to analyze the queue delay by adopting the standard M/G/1 queue model. In this model the input process is a Poisson process, and the service distribution is some general random process. In single level coding the time between services is a Geometrically distributed random variable. In order to use the M/G/1 model a crucial assumption on the system model must be made: the input arrives in blocks that have the same size as the transmission blocks. That is, the queue equation is normalized to blocks of transmission, where the block size is equal to an input block. The input process has a rate λ_{norm}, and at every embedding point the number of arrivals is measured in block units. This model is strongly limited by the constraint that arrival blocks are equal in size to transmitted blocks, since change of transmission rate means change in input block size. Therefore we do not adopt the M/G/1 queue model in this work.

A. *A simple example*

To gain some intuition on the zero-padding queue we use the following example. Assume that in equations (3) the input rate $\lambda/R < 1$ is fixed and deterministic, and the transmission rate is also fixed and deterministic, equals to 1. Denote the waiting time and queueing time in a zero-padding queue by w_n^{ZP} and q_n^{ZP} respectively. Let the queue be empty at $n = 0$, and for simplicity take R/λ and integer.

1) **Waiting time (normalized by λ)**: $\{w_0^{ZP}, w_1^{ZP}, ..., w_n^{ZP}, ...\} = \{0, 0, ..., 0, ...\}$.
2) **Queueing time (normalized by λ)**: $\{q_0^{ZP}, q_1^{ZP}, ..., q_n^{ZP}, ...\} = \{0, 1/R, ..., 1/R, ...\}$, as the only delay in the system is the transmission delay.

Hence **average** values of waiting time and queueing time are given by

1) **Average waiting time (normalized by λ)**: $\overline{w}^{ZP} = 0$.
2) **Average queueing time (normalized by λ)**: $\overline{q}^{ZP} = 1/R$ is the overall average delay in steady state.

B. *Steady-state conditions*

It is well known [46, Ch. 9] that the zero-padding queue is stable when the average input rate is strictly smaller than the service rate, and that the waiting time random process converges in distribution. Recall the Lindley equation in (3), given in a different form

$$w_{n+1} = \max(w_n + x_n, 0) \tag{6}$$

where $x_n \triangleq \lambda_{n+1} - R_n$ is a random variable of the difference process of the input and output random processes. With initial conditions w_1 (6) can also be expressed as,

$$w_{n+1} = \max(w_1 + x_1 + ... + x_n, x_2 + ... + x_n, ..., x_n, 0). \tag{7}$$

For each fixed n (7) shows that w_{n+1} depends on the partial sums of $x_1, ..., x_n$, summed in reverse order (aside from the initial condition w_1). Using standard tools from fluctuation theory (also known as theory of random walks), it is shown in [46, Ch. 9 Theorem 8] that,

Theorem 3.1: (stability) For every initial value of w_1, $\{w_n\}$ for an associated queue converges in distribution to w. For $E(x) < 0$ ($E\lambda < ER$), w is proper (i.e. it has zero mass at $\pm\infty :\ P(|w| < \infty) = 1$). For $E(x) \geq 0\ w = \infty$.

It may be concluded from theorem 3.1, that as long as the average input rate λ is strictly smaller than the average transmission rate, the system is stable. That is, the queue size will not grow unbounded (to ∞). Moreover, a steady state exists, as the queue size converges in distribution to a random variable, for which the mean value and its higher moments can be computed.

IV. Average Delay - Immediate Bounds

In this section average delay bounds are derived for the zero-padding queue. The upper and lower bounds based on known results for the G/G/1 queue. It is later shown that the delay upper bound is a rather tight one for multi-level coding, whereas the lower bound is quite loose. Therefore tighter lower bounds are derived for multi-level coding.

A. Upper bound

The queue in (3) may be expressed in a slightly different form,

$$w_{n+1} - y_n = w_n + x_n, \tag{8}$$

where y_n is the idle time process, which represents the amount of data that could have been served. When $n \to \infty$ it can be noticed that

$$Ey_\infty = -Ex = ER - E\lambda. \tag{9}$$

We now use the moment inequality. For an arbitrary non-negative random variable G

$$E[(G-v)_+^\alpha] \geq \frac{E(G^\alpha)}{(EG)^\alpha}(E[(G-v)_+])^\alpha, \quad for \;\; v \geq 0 \; and \; \alpha \geq 1. \tag{10}$$

An analytic derivation of the moment inequality (10) is presented by Daley [48]. The moment inequality for $\alpha = 2$ and $y_\infty = (w + \lambda - R)_- = (R - S)_+$ reduces after some algebra [46, Ch. 11-2] into

$$E[y_\infty^2] \geq (1 - \frac{E\lambda}{ER})^2 E(R^2). \tag{11}$$

By using (8), (9) and (11),

$$Ew \leq (E(x^2) - (1 - \frac{E\lambda}{ER})^2 E(R^2))/(2E(-x)) \tag{12}$$

which simplifies by equality to

$$Ew \leq \frac{\sigma_R^2 + \sigma_\lambda^2}{2(ER - E\lambda)} - (1 - \frac{E\lambda}{ER})\frac{\sigma_R^2}{2ER} \triangleq W_{UB,W}, \tag{13}$$

where σ_R^2 and σ_λ^2 are the variances of R and λ respectively, i.e. $\sigma_R^2 \triangleq ER^2 - (ER)^2$.

B. Lower bound

The lower bound derived here is based on tail properties of the queue output distribution. From equation (8) the average waiting time is given by [47, Ch. 2.3]

$$Ew = \frac{\sigma_R^2 + \sigma_\lambda^2}{2(ER - E\lambda)} + (ER - E\lambda)/2 - \frac{E(I^2)}{(EI)^2}, \tag{14}$$

where $I \sim (y_\infty | y_\infty \geq 0)$. The above can be lower bounded when R has the following properties. It is said that a random variable R has bounded mean residual life by γ (BMRL-γ) when

$$E(R - t | R > t) = \frac{\int\limits_t^\infty F_R^c(\tau) d\tau}{F_R^c(t)} \leq \gamma \quad \text{for all} \quad t > 0, \tag{15}$$

where $F_R^c(\tau)$ is the complementary CDF of R. If R has BMRL-γ, then [47, Ch. 2.3 eq (2.47)]

$$\frac{E(I^2)}{(EI)^2} \leq \gamma. \tag{16}$$

Hence the following is a lower bound for BMRL-γ transmission rate random variable R

$$Ew \geq \frac{\sigma_R^2 + \sigma_\lambda^2}{2(ER - E\lambda)} - (ER + E\lambda)/2 \triangleq W_{LB,K}. \tag{17}$$

where $\gamma = ER$.

V. Queueing and Multi-Layer Coding

In this section the transmitter uses superposition coding for single transmit antenna, and one or more receive antennas. Upper and lower delay bounds are derived for two level code layering. The derivation relies on the relationship of the Laplace transform and the waiting time cumulative distribution function (CDF). Similar type of bounds were derived for the outage approach in [10]. We hereby derive tight bounds for finite level coding and continuous layering (broadcasting).

A. Maximal Throughput in Two-level layering

A two-level code layering for the SISO channel was presented by Liu *et. al.* [49]. Extensions for multiple-input single-output (MISO) setting and single-input multiple-output (SIMO) setting are presented in [32]. In a two level coding, as in the broadcasting, the receiver views a degraded broadcast channel. Consider a two layer code of rates R_1 and R_2, such that the transmission rate $R = R_1 + R_2$. Two channel fading power parameters s_1 and s_2 are defined respectively. We restrict $0 \le s_1 \le s_2$ without loss of generality. When fading parameter $0 \le s$ no layer can be decoded. When $s_1 \le s \le s_2$ the first layer **only** can be decoded while treating the other as interference. When $s_2 \le s$ both layers can be decoded by initially decoding R_1, cancelling it from the received signal and then decoding of R_2, in better signal to interference ratio (SIR) conditions. The fixed rate of the first layer is

$$R_1 = \log\left(1 + \frac{(1-\beta)Ps_1}{1+\beta Ps_1}\right), \tag{18}$$

where the power allocated to layers R_1 and R_2 is $(1-\beta)P$ and βP respectively. The second layer can be decoded for a fading parameter $s \ge s_2$. Obviously, in this case the first layer can be decoded prior to R_2. Therefore, the rate R_2 has no inter-layer interference and its fixed rate is given by

$$R_2 = \log\left(1 + \beta P s_2\right). \tag{19}$$

The expression of the maximal achievable average rate is given by

$$R_{2L} = \max_{\beta, s_1, s_2} P_{succ}(s_1)R_1 + P_{succ}(s_2)R_2, \tag{20}$$

where $P_{succ}(s_i)$ is the same probability of successful decoding of layer i. Optimal β for which R_{2L} achieves maximum is specified by [49]

$$\beta_{opt} = \Phi\left(\frac{s_2 P_{succ}(s_2) - s_1 P_{succ}(s_1)}{P s_1 s_2 (P_{succ}(s_1) - P_{succ}(s_2))}\right) \tag{21}$$

where

$$\Phi(x) \triangleq \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

The maximal achievable rate can be further optimized over $s_1,\ s_2$ numerically. Note that the difference between the MISO and SIMO settings lies in (18), (19) where the power is normalized by the number of transmit antennas. That is P in (18) and (19) is replaced by $\frac{P}{M}$ in a MISO setting [33].

B. Delay bounds in Two-level layering

In this section bounds that exploit the multi-level coding queue equation are derived. The number of layers is restricted here to two level layering. It is later relaxed and extended to the derivation of multiple-level coding and continuous coding bounds. The queue size equation (3) is also well known as the Lindley equation [46]. It describes the queue size at the beginning of every time slot (embedding point). A two level coding queue size equation is specified hereby,

$$W_{n+1} = \begin{cases} W_n + X_n & W_n + X_n \geq 0 \\ 0 & W_n + X_n < 0 \end{cases} \tag{22}$$

where

$$X_n \triangleq \lambda - \nu_{1,n} R_1 - \nu_{2,n} R_2 \tag{23}$$

and we have assumed that the input rate is deterministic λ, which means that at every embedding point a new input block λ arrives. Layering rates R_1 and R_2 are specified by (18) and (19) respectively. The outage region for layering is described by $\nu_{1,n},\ \nu_{2,n}$. The random variables can be defined by the channel realization thresholds as follows

$$\nu_{1,n} = \begin{cases} 1 & s_1 \leq s_n \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

$$\nu_{2,n} = \begin{cases} 1 & s_2 \leq s_n \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

where s_n is the fading power realization at the n^{th} time-slot. Calculating the CDF of the queue size at these embedding points will enable the calculation of the CDF at every time instant. A recursive notion of the CDF $F_W(w)$ of the queue size [50, Ch. 8]

$$F_W(w) = \begin{cases} 0 & w < 0 \\ \int\limits_{-\infty}^{w} F_W(w-\tau)dF_X(\tau) & w \geq 0 \end{cases} . \tag{26}$$

In our setting the probability density $dF_X(\tau)$ of X (23) is

$$dF_X(x) = p_1\delta(x-(\lambda-R_1-R_2)) + p_2\delta(x-(\lambda-R_1)) + \overline{p}\delta(x-\lambda), \tag{27}$$

where $p_1 = Prob\{s_n \geq s_2\}$, $p_2 = Prob\{s_1 \leq s_n \geq s_2\}$ and $\overline{p} = 1 - p_1 - p_2$.

Theorem 5.1: Queue average size and average delay for two level code layering are upper and lower bounded by

$$EW_2 \geq \frac{(R_1+R_2)\lambda(1-p_1) - p_2R_1(\lambda+R_2)}{2(p_1(R_1+R_2)+p_2R_1-\lambda)} \tag{28}$$

$$EW_2 \leq \frac{p_1R_2^2 + 2p_1R_2(R_1-\lambda) + (p_2+p_1)R_1^2 - 2\lambda R_1(p_1+p_2) + \lambda^2}{2(p_1(R_1+R_2)+p_2R_1-\lambda)} \tag{29}$$

and the average delay normalized by the input rate λ is bounded by

$$EW_{2,\lambda} \geq \frac{(R_1+R_2)\lambda(1-p_1) - p_2R_1(\lambda+R_2)}{2\lambda(p_1(R_1+R_2)+p_2R_1-\lambda)} \tag{30}$$

$$EW_{2,\lambda} \leq \frac{p_1R_2^2 + 2p_1R_2(R_1-\lambda) + (p_2+p_1)R_1^2 - 2\lambda R_1(p_1+p_2) + \lambda^2}{2\lambda(p_1(R_1+R_2)+p_2R_1-\lambda)} \tag{31}$$

Proof: See *Appendix A.*

Similarly the upper bound in (13) can be explicitly specified for the two-level code layering approach. To obtain σ_R^2 the following is required,

$$\sigma_{R_{2L}}^2 \triangleq p_2R_1^2 + p_1(R_1+R_2)^2 - R_{2L,av}^2 \tag{32}$$

where

$$R_{2L,av} \triangleq p_1(R_1+R_2) + p_2R_1 \tag{33}$$

Thus we have the following result,

Corollary 5.1: Queue average size and average delay for a two-level code layering are upper bounded (13) by

$$EW_{2L} \leq \frac{\sigma^2_{R_{2L}}}{2(R_{2L,av} - \lambda)} - (1 - \frac{\lambda}{R_{2L,av}})\frac{\sigma^2_{R_{2L}}}{2R_{2L,av}}, \tag{34}$$

and the average delay normalized by the input rate λ is upper bounded by

$$EW_{\lambda,2L} \leq \frac{\sigma^2_{R_{2L}}}{2\lambda(R_{2L,av} - \lambda)} - (1 - \frac{\lambda}{R_{2L,av}})\frac{\sigma^2_{R_{2L}}}{2R_{2L,av}\lambda}, \tag{35}$$

where $\sigma^2_{R_{2L}}$ and $R_{2L,av}$ are given by (32) and (33) respectively.

VI. Delay bounds for Finite Level Code Layering

In this section the Lindley equation [46] for finite-level code layering is introduced, for some number of code layers K. The Lindley equation describes the queue size at the beginning of every time slot (embedding point). For finite level coding at the queue output we have, like in (22),

$$W_{n+1} = \begin{cases} W_n + X_n & W_n + X_n \geq 0 \\ 0 & W_n + X_n < 0 \end{cases} \tag{36}$$

where the queue update random variable X_n depends on the number of layers in the code. Its realization specifies the difference between the number of layers successfully decoded and the queue input λ,

$$X_n \triangleq \lambda - \sum_{i=1}^{K} \nu_{i,n} R_i \tag{37}$$

and we have assumed that the input rate is deterministic λ and so are the layering rates $\{R_i\}_{i=1}^{K}$. The outage region for layering is determined by $\{\nu_{i,n}\}_{i=1}^{K}$. The associated fading power thresholds are denoted $\{s_{th,i}\}_{i=1}^{K}$. The random variables (RV) $\{\nu_{i,n}\}_{i=1}^{K}$ are related to the fading thresholds as follows

$$\nu_{i,n} = \begin{cases} 1 & s_{th,i} \leq s_n \leq s_{th,i+1} \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

where s_n is the fading power realization at the n^{th} time-slot, and $s_{th,K+1} = \infty$. Every RV $\nu_{i,n}$ has a probability denoted p_{K-i+1} for being 1. Note that outage probability $\overline{p} = 1 - \sum\limits_{i=1}^{K} p_i$, where $\overline{p}$ stands for the probability that no layer is decoded. Calculating the CDF of the queue size at these embedding points will enable the calculation of the CDF at every time instant. In equivalence to two level layering a recursive notion of the CDF $F_W(w)$ of the queue size (26) can be used. In this setting the probability density $dF_X(\tau)$ of X (37) is

$$dF_X(x) = \sum_{i=1}^{K} p_i \delta(x - (\lambda - \sum_{j=1}^{K-i+1} R_j)) + \overline{p}\delta(x - \lambda) \tag{39}$$

where $p_i = Prob\{s_{th,i} \leq s_n \leq s_{th,i+1}\}$ for $i = 1, ..., K$ and $s_{th,K+1} = \infty$.

Theorem 6.1: Queue average size and average delay for K-level code layering are upper and lower bounded by

$$EW_K \geq \frac{(\Re_K - \lambda)(\sum\limits_{i=1}^{K} p_i \Re_{K-i+1} - \lambda) - (\Re_K - \lambda)^2 + \sum\limits_{i=1}^{K} p_i(\Re_K - \Re_{K-i+1})^2 + \overline{p}\Re_K^2}{2(\sum\limits_{i=1}^{K} p_i \Re_{K-i+1} - \lambda)} \tag{40}$$

$$EW_K \leq \frac{2(\Re_K - \lambda)(\sum\limits_{i=1}^{K} p_i \Re_{K-i+1} - \lambda) - (\Re_K - \lambda)^2 + \sum\limits_{i=1}^{K} p_i(\Re_K - \Re_{K-i+1})^2 + \overline{p}\Re_K^2}{2(\sum\limits_{i=1}^{K} p_i \Re_{K-i+1} - \lambda)} \tag{41}$$

where $\Re_V \triangleq \sum_{j=1}^{V} R_j$, and the average delay normalized by the input rate λ is bounded by

$$EW_{K,\lambda} \geq \frac{(\Re_K - \lambda)(\sum\limits_{i=1}^{K} p_i \Re_{K-i+1} - \lambda) - (\Re_K - \lambda)^2 + \sum\limits_{i=1}^{K} p_i(\Re_K - \Re_{K-i+1})^2 + \overline{p}\Re_K^2}{2\lambda(\sum\limits_{i=1}^{K} p_i \Re_{K-i+1} - \lambda)} \tag{42}$$

$$EW_{K,\lambda} \leq \frac{2(\Re_K - \lambda)(\sum\limits_{i=1}^{K} p_i \Re_{K-i+1} - \lambda) - (\Re_K - \lambda)^2 + \sum\limits_{i=1}^{K} p_i(\Re_K - \Re_{K-i+1})^2 + \overline{p}\Re_K^2}{2\lambda(\sum\limits_{i=1}^{K} p_i \Re_{K-i+1} - \lambda)} \tag{43}$$

Proof: See *Appendix B.*

The upper bound in (13) can be explicitly specified for the finite-level code layering approach. To obtain σ_R^2 the following is required,

$$\sigma^2_{R_{KL}} \;\; \triangleq \;\; \sum_{i=1}^{K} p_i \Re^2_{K-i+1} - (R_{KL,av})^2 \tag{44}$$

where

$$R_{KL,av} \triangleq \sum_{i=1}^{K} p_i \Re_{K-i+1}. \tag{45}$$

Corollary 6.1: Queue average size and average delay for a K-level code layering are upper bounded (13) by

$$EW_{KL} \leq \frac{\sigma^2_{R_{KL}}}{2(R_{KL,av} - \lambda)} - (1 - \frac{\lambda}{R_{KL,av}})\frac{\sigma^2_{R_{KL}}}{2R_{KL,av}}, \tag{46}$$

and the average delay normalized by the input rate λ is upper bounded by

$$EW_{\lambda,KL} \leq \frac{\sigma^2_{R_{KL}}}{2\lambda(R_{KL,av} - \lambda)} - (1 - \frac{\lambda}{R_{KL,av}})\frac{\sigma^2_{R_{KL}}}{2R_{KL,av}\lambda}, \tag{47}$$

where $\sigma^2_{R_{KL}}$ and $R_{KL,av}$ are given by (44) and (45) respectively.

VII. DELAY BOUNDS FOR CONTINUUM BROADCASTING

We adhere to the broadcasting approach for a SISO channel [31]. In this approach the transmitter also sends multi-layer coded data. The receiver decodes the maximal number of layers given a channel realization (per-block). However, as opposed to finite-level code layering, here the layering may be a continuous function of the channel fading parameter. That is, the number of layers is not limited in advance, and an incremental rate with a differential power allocation is associated with every layer. The differential rate per layer is given by

$$dR(s) = \log\left(1 + \frac{s\rho(s)ds}{1 + sI(s)}\right) = \frac{s\rho(s)ds}{1 + sI(s)} \tag{48}$$

where $\rho(s)ds$ is the transmit power of a layer parameterized by s, intended for receiver s, which also designates the transmit power distribution. The right hand-side equality is justified in [51]. Information streams intended for receivers indexed by $u > s$ are undetectable and play a role of additional interfering noise, denoted by $I(s)$. The interference for a fading power s is

$$I(s) = \int\limits_s^\infty \rho(u)du, \tag{49}$$

which is also a monotonically decreasing function of s. The total transmitted power is the overall collected power assigned to all layers,

$$P = \int\limits_0^\infty \rho(u)du = I(0). \tag{50}$$

As mentioned earlier, the total achievable rate for a fading realization s is an integration of the fractional rates over all receivers with successful layer decoding capability,

$$R(s) = \int_0^s \frac{u\rho(u)du}{1 + uI(u)}. \tag{51}$$

Average rate is achieved with sufficiently many transmission blocks, each viewing an independent fading realization. Therefore, the total average rate R_{bs} over all fading realizations is

$$R_{bs} = \int\limits_0^\infty du\ f(u)R(u) = \int_0^\infty du(1-F(u))\frac{u\rho(u)}{1+uI(u)} \tag{52}$$

where $f(u)$ is the probability distribution function (PDF) of the fading power, and $F(u) = \int\limits_0^u da f(a)$ is the corresponding cumulative distribution function (CDF).

It is possible extend the finite-level code layering bounds derived above to this broadcast setting. The bounds in Eq. (40) and (41) could be used for broadcasting after performing the following modifications:

1) The number of layers is unlimited, that is $K \to \infty$.
2) Every layer i is associated with a fading parameter s, hence the layering is continuous. Every Rate R_i is associated now with a differential rate $dR(s)$ (48).
3) The cumulative rate $\Re_K$ should be replaced by

$$R_T = \int\limits_0^\infty dR(s)ds. \tag{53}$$

4) The sum $\sum\limits_{i=1}^{K} p_i \Re_{K-i+1}$ is actually the average rate and it turns to be R_{bs} (52) in the continuum case.
5) Finally, in finite level coding the expression $\sum\limits_{i=1}^{K} p_i(\Re_K - \Re_{K-i+1})^2 + \overline{p}\Re_K^2$ turns out to be

$$\begin{aligned} R_{d,bs}^2 &\triangleq \int\limits_0^\infty du f(u)\left[R_T - \int\limits_0^u dR(s)\right]^2 \\ &= \int\limits_0^\infty du f(u)\left[\int\limits_u^\infty dR(s)\right]^2 \\ &= 2\int\limits_0^\infty du F(u) dR(u) \int\limits_u^\infty dR(s) \end{aligned} \tag{54}$$

in the continuous case, where $dR(u)$ and $R(u)$ are specified in (48) and (51) respectively.

Corollary 7.1: Queue average size and average delay for a continuous code layering are upper and lower bounded by

$$EW_{bs} \geq \frac{R_T - \lambda}{2} + \frac{R_{d,bs}^2 - (R_T-\lambda)^2}{2(R_{bs}-\lambda)} \tag{55}$$

$$EW_{bs} \leq (R_T - \lambda) + \frac{R_{d,bs}^2 - (R_T-\lambda)^2}{2(R_{bs}-\lambda)} \tag{56}$$

and the average delay normalized by the input rate λ is bounded by

$$EW_{\lambda,bs} \geq \frac{R_T - \lambda}{2\lambda} + \frac{R_{d,bs}^2 - (R_T - \lambda)^2}{2\lambda(R_{bs} - \lambda)} \tag{57}$$

$$EW_{\lambda,bs} \leq \frac{R_T - \lambda}{\lambda} + \frac{R_{d,bs}^2 - (R_T - \lambda)^2}{2\lambda(R_{bs} - \lambda)} \tag{58}$$

where R_{bs}, R_T and $R_{d,bs}^2$ are specified in (52), (53) and (54) respectively.

Similarly the upper bound in (13) can be explicitly specified for the continuous code layering approach. To obtain σ_R^2 the following is required,

$$\begin{aligned}\sigma_{R_{bs}}^2 &\triangleq \int_0^\infty du f(u) \left[R(u)\right]^2 - R_{bs}^2 \\ &= \int_0^\infty du f(u) \left[\int_0^u dR(s)\right]^2 - R_{bs}^2 \\ &= 2\int_0^\infty du(1 - F(u))dR(u) \int_0^u dR(s) - R_{bs}^2 \\ &= 2\int_0^\infty du(1 - F(u))dR(u)R(u) - R_{bs}^2.\end{aligned} \tag{59}$$

Corollary 7.2: Queue average size and average delay for a continuous code layering are upper bounded (13) by

$$EW_{bs} \leq \frac{\sigma_{R_{bs}}^2}{2(R_{bs} - \lambda)} - (1 - \frac{\lambda}{R_{bs}})\frac{\sigma_{R_{bs}}^2}{2R_{bs}}, \tag{60}$$

and the average delay normalized by the input rate λ is upper bounded by

$$EW_{\lambda,bs} \leq \frac{\sigma_{R_{bs}}^2}{2\lambda(R_{bs} - \lambda)} - (1 - \frac{\lambda}{R_{bs}})\frac{\sigma_{R_{bs}}^2}{2R_{bs}\lambda}, \tag{61}$$

where R_{bs} and $\sigma_{R_{bs}}^2$ are given by (52) and (59) respectively.

Minimizing the average delay in the continuous case requires finding the optimal power distribution $\rho(s)$ (49). As in the case of finite level coding the optimization problem of finding the optimal power allocation does not lend itself to an analytic solution. Numerical optimization is impossible here, as opposed to the finite level case, where the number of optimization variables is small. Here the function subject to optimization is continuous. The target functional in the optimization problem underhand for continuous layering does not have a localization property [52]. A functional with localization property can be written as an integral(s) of some target function. Our functional contains a ratio of integrals and further multiplication of integrals, which cannot be converted to an integral(s) over a single target function. This type of functional is also denoted as a nonlocal functional in Gelfand et. al. [52]. In such cases it is preferable to look for an approximate representation, of the nonlocal functional, which has the localization property. Alternatively, approximate target functions with reduced degrees of freedom may be optimized.

In order to reduce degrees of freedom, and introduce a tractable optimization problem, a power distribution $\rho(s)$ is selected in advance, while inserting two unconstrained parameters, and evaluating the delay numerically. This provides an approximation of the optimal continuous layering delay, while keeping in mind that the delay performance could be further minimized, if an optimal power distribution was known.

The selected subject power distribution $\rho(s)$ is based on the maximal throughput realizing function $\rho(s)$. As already known [30], for the Rayleigh fading channel, with a fading power distribution $f_s(u) = e^{-u}$, the throughput optimal interference power distribution is given by [30]

$$I(s) = \frac{1}{s^2} - \frac{1}{s} \qquad s_0 \leq s \leq s_1 \tag{62}$$

where $I(s_0) = P$ and $I(s_1) = 0$, and $\rho(s) = -\frac{dI(s)}{ds}$. When using (62) for computation of average delay, the delay is much higher than that of optimal (minimal) delay in finite level coding for a large range of input rates λ. Thus the approximate interference distribution chosen is as follows

$$I(s) = \frac{c_0}{s^2} - \frac{c_1}{s} \qquad s_0 \leq s \leq s_1 \tag{63}$$

where c_0 and c_1 are fixed scalar coefficients ($c_0 \geq 0$ and $c_1 > 0$), chosen as to minimize the delay for every input rate λ. The relations $I(s_0) = P$ and $I(s_1) = 0$ can be specified by using (63)

$$s_1 = \frac{c_0}{c_1} \tag{64}$$

and

$$s_0 = \frac{-c_0 + \sqrt{c_0^2 + 4c_1 P}}{2P}. \tag{65}$$

Having defined this, average delay upper and lower bounds may be computed for various power distributions given in (63). The upper bound in (61) is minimized numerically over c_0 and c_1, for every input rate λ. In general, the resulting minimal average delay is still an upper bound on the global minimal delay, since the power distribution function is only an approximation of the optimal function, based on the corresponding maximal throughput achieving one.

VIII. Minimal average delay - numerical results

A. ***Outage approach*** *delay bounds comparison*

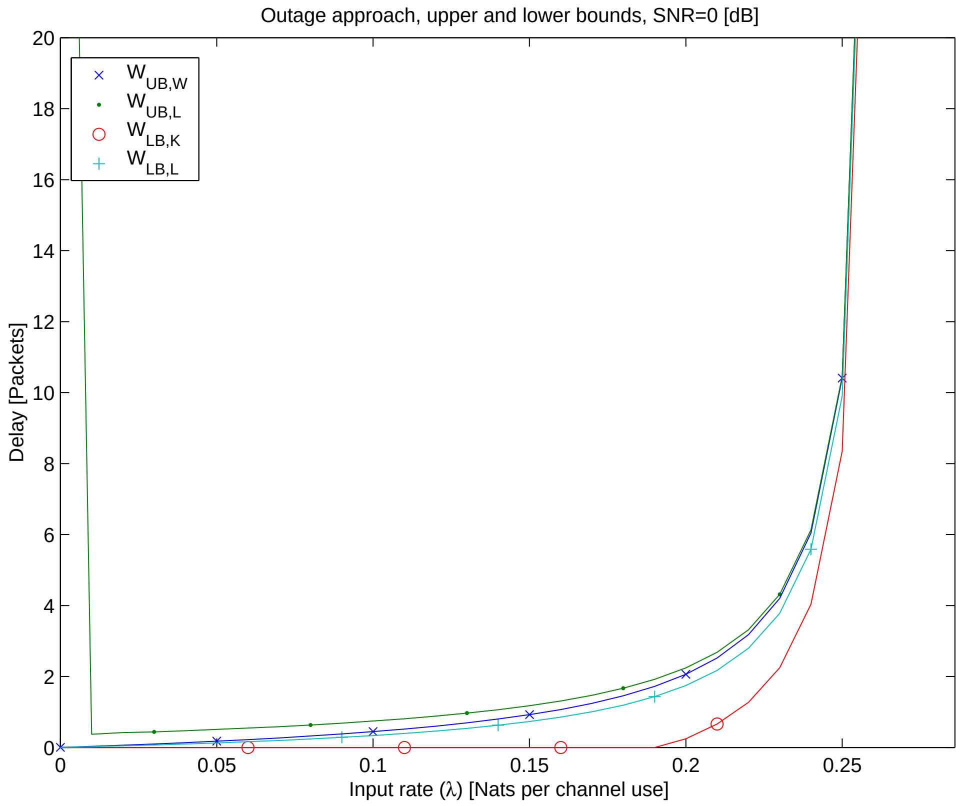

Fig. 2. Average delay - outage approach. Delay is demonstrated for minimal delay power assignment (SNR=0dB) Bounds of [10] are compared with $W_{UB,W}$ (13) and $W_{LB,K}$ (17).

Figure 2, 3 demonstrate the average delay bounds for the outage approach. The rate and power allocation of the transmitter are optimized for every λ such that the upper bound $W_{UB,W}$ (13) is minimized. All other bounds are computed for the same fading parameter threshold as optimized for $W_{UB,W}$. It may be seen that $W_{UB,W}$ and the lower bound $W_{LB,L}$ from [10] closely predict the average delay, as these two bounds are tightest, even for low SNR. Furthermore, the lower bound $W_{LB,K}$ (17) is not tight, and therefore will not be used in following numerical results presentation.

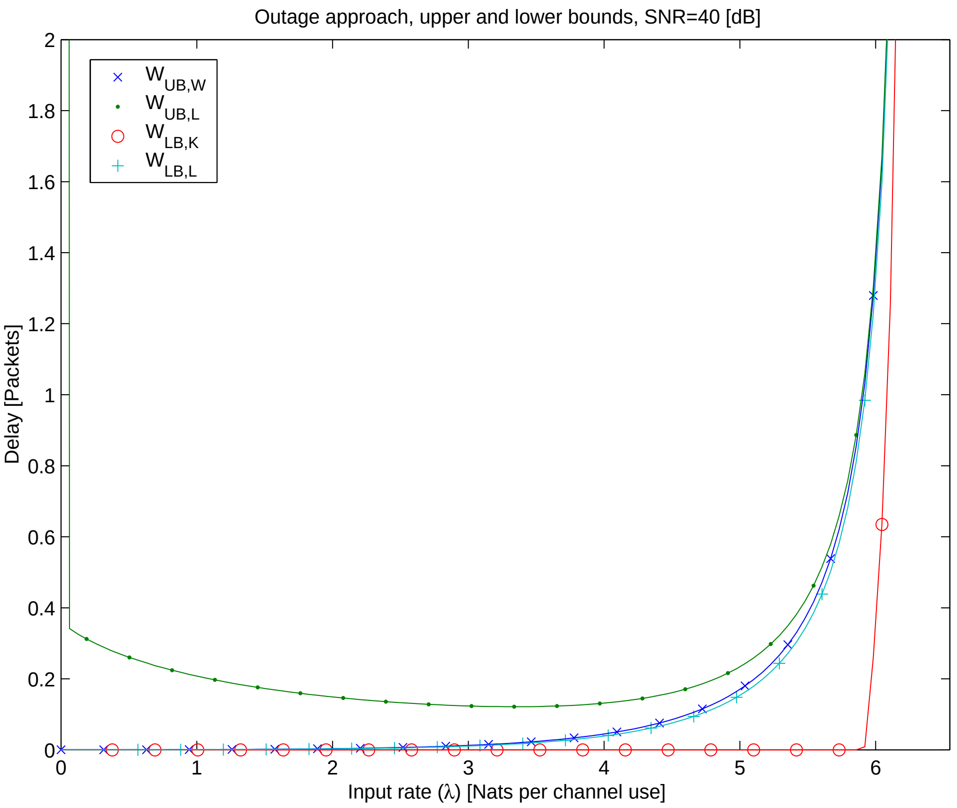

Fig. 3. Average delay - outage approach. Delay is demonstrated for minimal delay power assignment (SNR=40dB) Bounds of [10] are compared with $W_{UB,W}$ (13) and $W_{LB,K}$ (17).

B. ***Multi-level coding approach*** *delay bounds comparison*

Figures 4-5 demonstrate the average delay upper bounds for outage, two, three-level coding and continuous layering (broadcasting). The rate and power allocation for each approach are jointly optimized for every λ. The upper bound $W_{nL,UB,W}$ (13) is minimized for every coding approach over all power and rate allocation free parameters. In the continuous layering case the delay is optimized over the two variables in (63) to produce a minimal delay. In three level coding there are **three** fading power thresholds, and **two** power allocation fractions specifying together the rate allocation for each layer. $W_{nL,UB,W}$ is optimized over all these parameters, which are then used to compute the upper bounds $W_{nL,UB,L}$ (43) for each coding strategy. As may be noticed the $W_{nL,UB,W}$ (13) upper bounds are tighter than $W_{nL,UB,L}$ bounds, particularly in low input rates, and high SNRs. For this reason we use only $W_{nL,UB,W}$ (13) upper bounds for computation of average delays and comparison to $W_{nL,LB,L}$ lower bounds in the following.

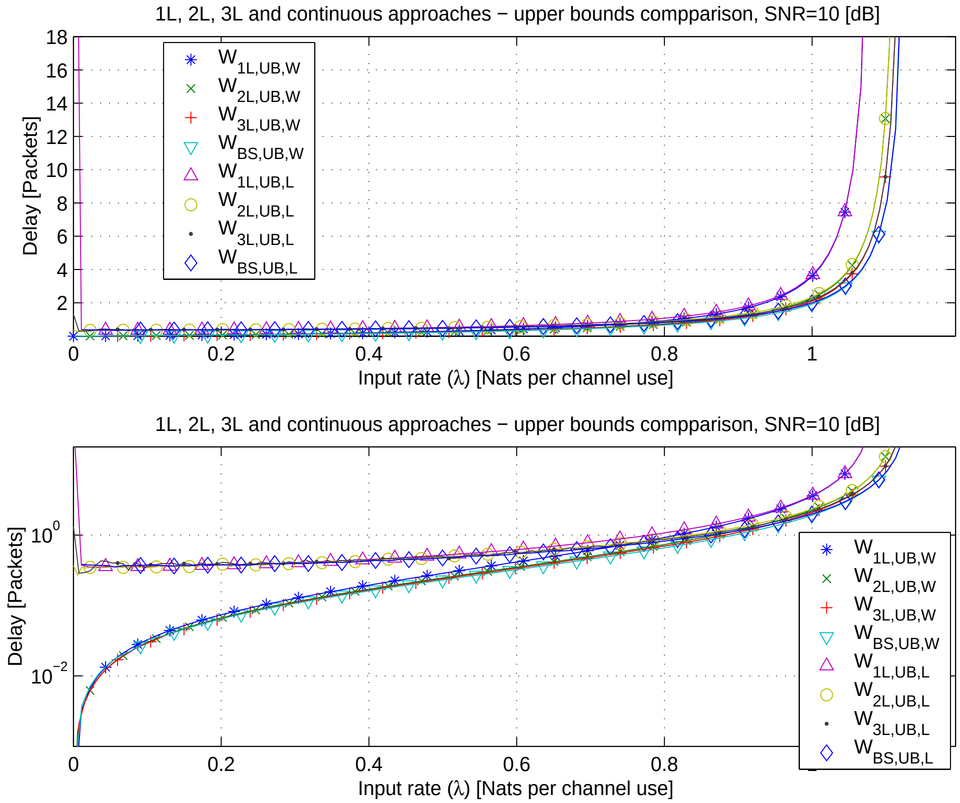

Fig. 4. Average delay for outage approach, 2-level and 3-level coding and continuous layering (BS), for SNR=10dB. The bounds $W_{nL,UB,L}$ denote the n-level coding upper bounds specified in (43). The bounds $W_{nL,UB,W}$ denote the n-level coding upper bounds specified in its general form in (13).

1) ***Fixed SNR, variable input rate*** *λ:* Figures 6 - 10 demonstrate the average delay bounds (upper and lower) for outage, two, three-level coding and continuous layering. The rate and power allocation for each approach is optimized for every λ. The upper bound $W_{nL,UB,W}$ (13) is minimized for every coding approach over all free parameters. The lower bounds $W_{nL,LB,L}$ (42) are then computed with the same power and rate parameters used for computing $W_{nL,UB,W}$. As may be noticed in low SNR multi-level coding does not show much improvement over single level coding (outage). However, in high SNRs and moderate input rates the three level coding has a pronounced delay improvement over the outage approach. Figures 9 - 10 also show, as expected, that delay gains in two level coding over outage are greater than those of three-level coding compared to two level coding. This suggests that in limit of continuous layering there will be no significant delay improvement over the three-level coding delays.

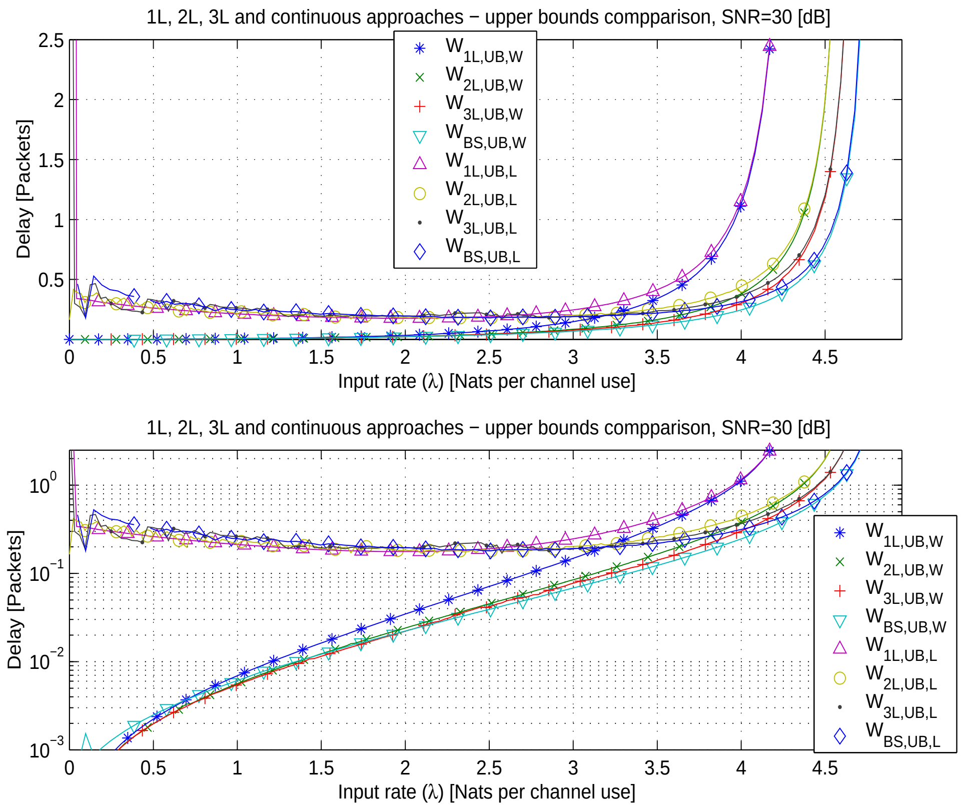

Fig. 5. Average delay for outage approach, 2-level and 3-level coding and continuous layering (BS), for SNR=30dB. The bounds $W_{nL,UB,L}$ denote the n-level coding upper bounds specified in (43). The bounds $W_{nL,UB,W}$ denote the n-level coding upper bounds specified in its general form in (13).

*2) **Fixed input rate λ, variable SNR**:* Figures 12 - 15 demonstrate the average delay bounds (upper and lower) for outage, two, three-level coding and continuous layering. The bounds are computed for some given input rates, and presented as function of the SNR. The rate and power allocation for each approach are jointly optimized for every SNR value. The upper bound $W_{nL,UB,W}$ (13) is minimized for every coding approach over all free parameters. In the continuous layering case the delay is optimized over the two variables in (63) to produce a minimal delay. Interestingly, as the SNR increases the delay ratio between the different methods is maintained.

From these figures the SNR gain for some input rate and expected packet delay may be computed. For example, in Figure 16 the input rate is $\lambda = 5$ Nats/channel use. For an expected packet delay of 0.2, an SNR of ∼39.5dB is required for single-level coding, ∼36.5dB is required for two-level coding, ∼35.5dB is required for three-level coding and ∼35dB is required for

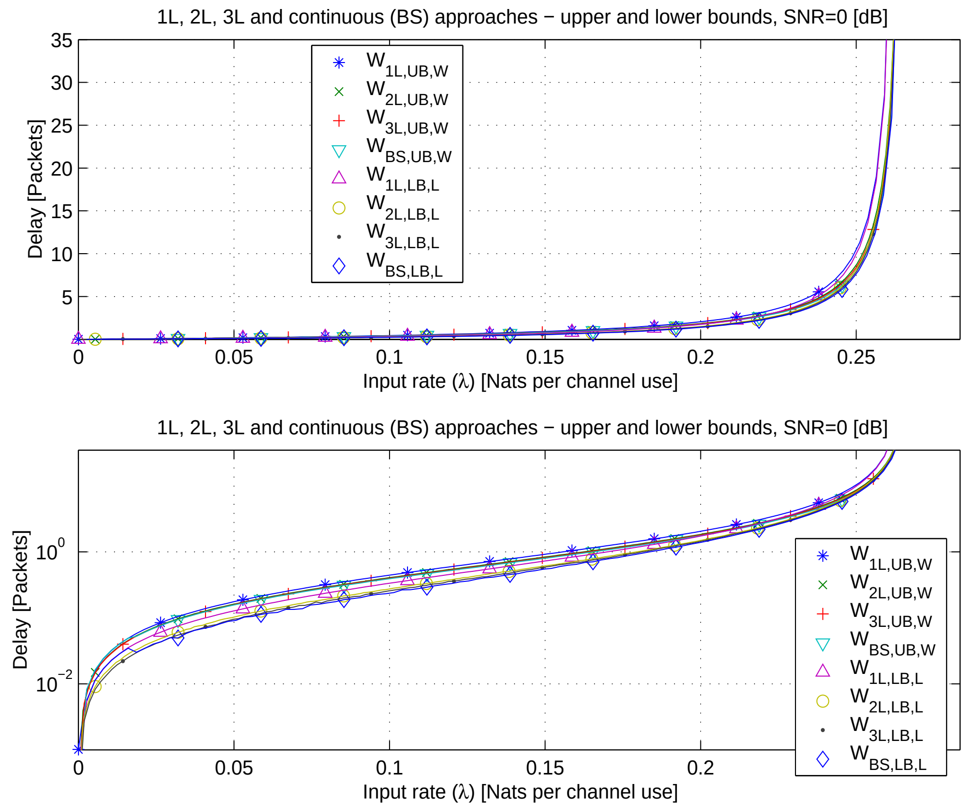

Fig. 6. Average delay for outage approach, 2-level and 3-level coding and continuous layering (BS), for SNR=0dB. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

continuous layering. This suggests a gain of 3dB of two level coding over outage, another 1dB for three level coding, and an additional 0.5dB for continuous layering. In total, continuous layering gains ∼4.5dB over outage approach in terms of average delay.

When comparing to Figure 11, throughput gain of two level coding over outage is ∼1.5dB at SNR=40dB, as opposed to a 3dB gain in delay performance. When comparing outage to continuous layering, the throughput gain is ∼3.3dB, whereas the delay gain is more than ∼4.5 dB. We have used an approximation of the power distribution (63) in continuous layering, therefore the maximal delay gain is expected to be even higher. This clearly shows that when considering delay as a performance measure, code layering may give pronounced gains in delay, which were not predicted when analyzing only the maximal throughput performance.

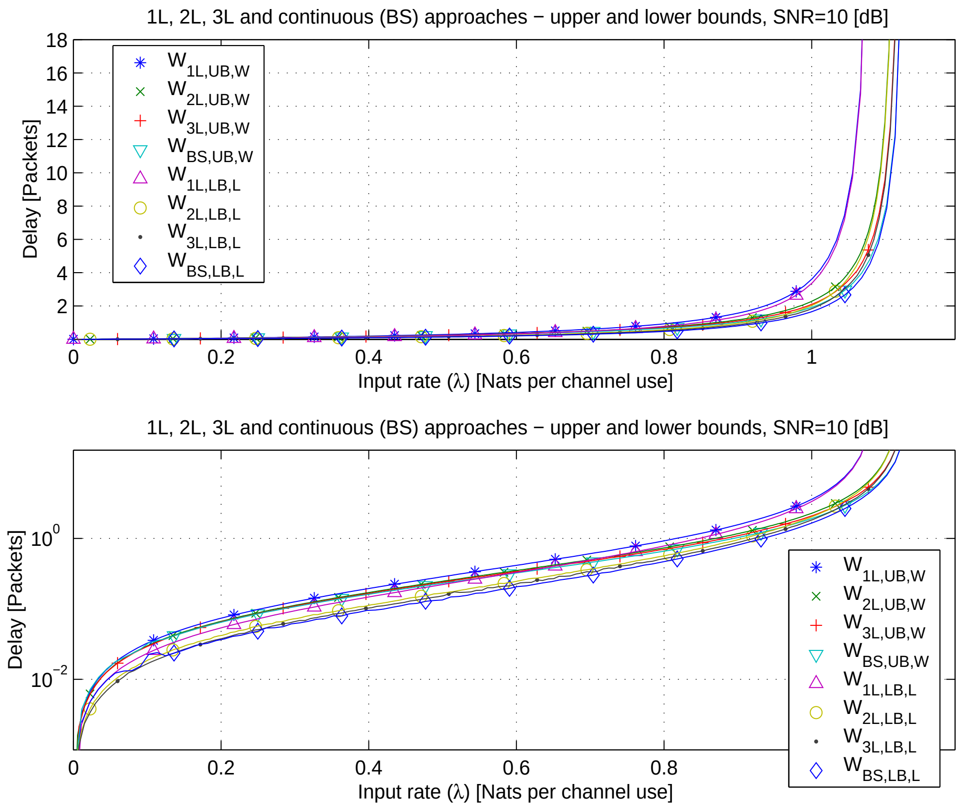

Fig. 7. Average delay for outage approach, 2-level and 3-level coding and continuous layering (BS), for SNR=10dB. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

IX. Conclusion

We have studied a single-server queue concatenated with a multi-level channel encoder. The main focus of this work is on minimization of the average delay of a packet from entering the queue until completion of successful service. Tight bounds are derived for the average delay for different numbers of coded layers. The bounds are optimized numerically for a Rayleigh block fading channel.

Delay bounds are also derived for continuous layering (single user broadcast approach). The optimizing power distribution of the minimal delay is approximated, and numerically evaluated.

An interesting observation from the numerical results is that when considering delay as a performance measure, code layering may give pronounced performance gains in terms of delay, which are more impressive than those associated with throughput. This makes layering more

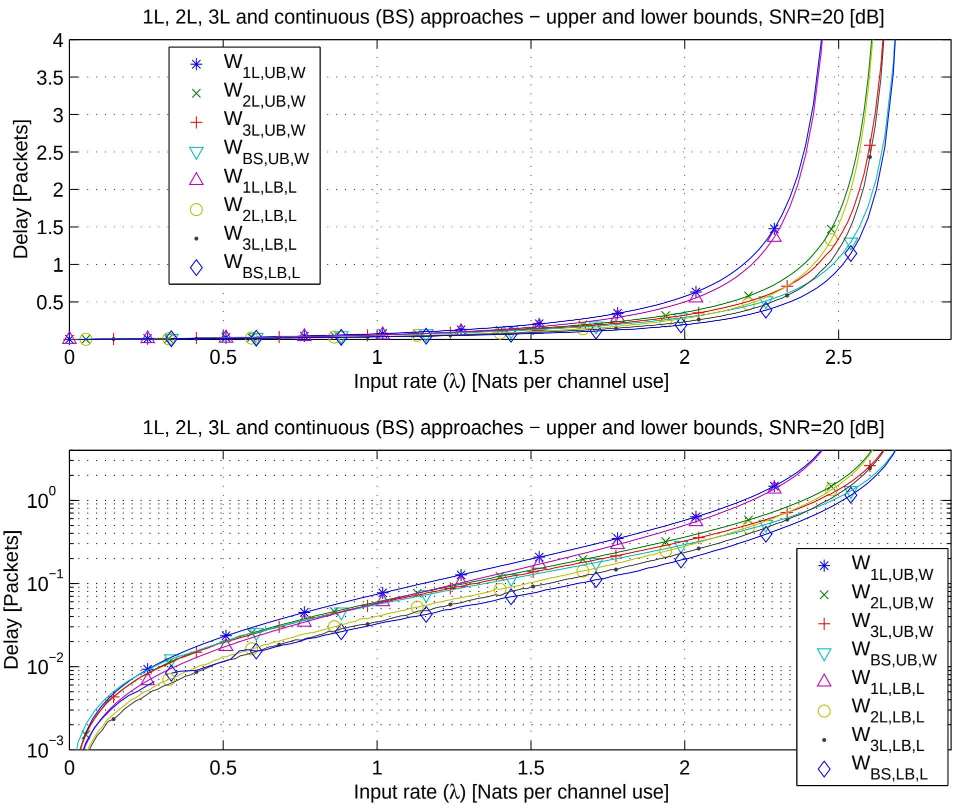

Fig. 8. Average delay for outage approach, 2-level and 3-level coding and continuous layering (BS), for SNR=20dB. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

attractive when communicating under stringent delay constraints.

The results can be extended to SIMO, and MISO systems in a direct manner, and the derived bounds may be used just by replacing the fading parameter distribution with the one corresponding to the multiple antenna case.

Appendix A

Proof of Theorem 5.1

Proof: The proof consists of two parts. In the first part we assume that $R_1 \leq \lambda$ and in the second part the inverse is assumed $R_1 > \lambda$. It is shown that the same bounds (28)-(29) are reached in both cases.

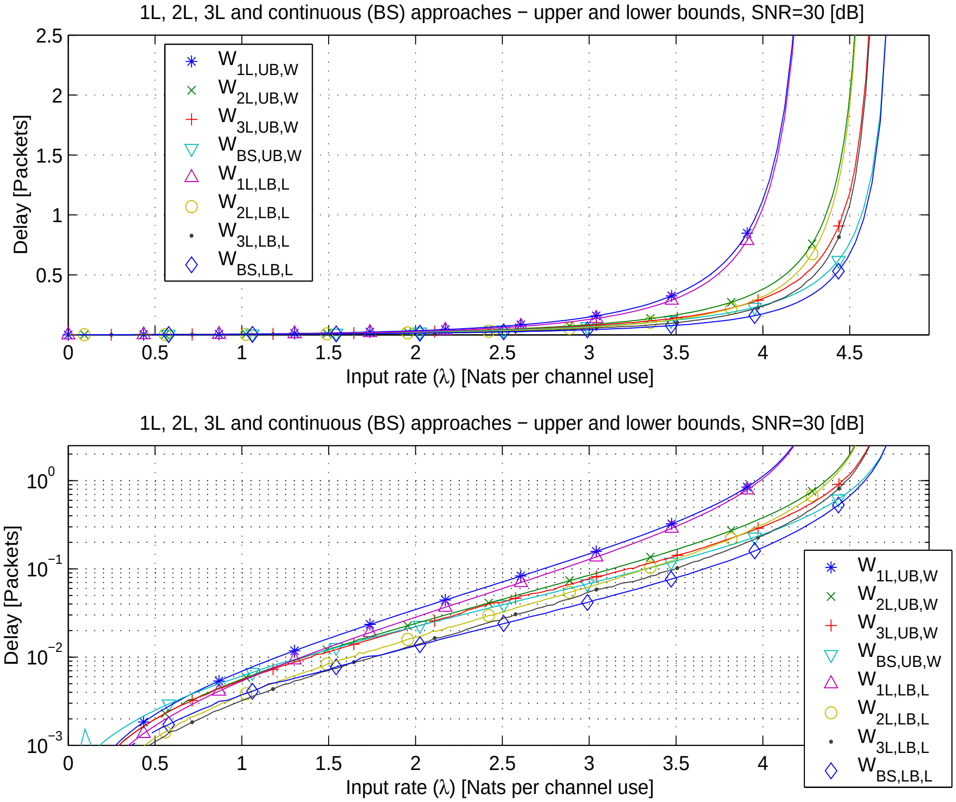

Fig. 9. Average delay for outage approach, 2-level and 3-level coding and continuous layering (BS), for SNR=30dB. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

A) $R_1 \le \lambda$**:** The queue size CDF follows from (26) and (27),

$$F_W(w) = \begin{cases} 0 & w < 0 \\ p_1 F_W(w - (\lambda - R_1 - R_2)) & 0 \le w \le \lambda - R_1 \\ p_1 F_W(w - (\lambda - R_1 - R_2)) + p_2 F_W(w - (\lambda - R_1)) & \lambda - R_1 \le w \le \lambda \\ p_1 F_W(w - (\lambda - R_1 - R_2)) + p_2 F_W(w - (\lambda - R_1)) + \overline{p} F_W(w - \lambda) & \lambda \le w \end{cases} \quad \textbf{(A.1)}$$

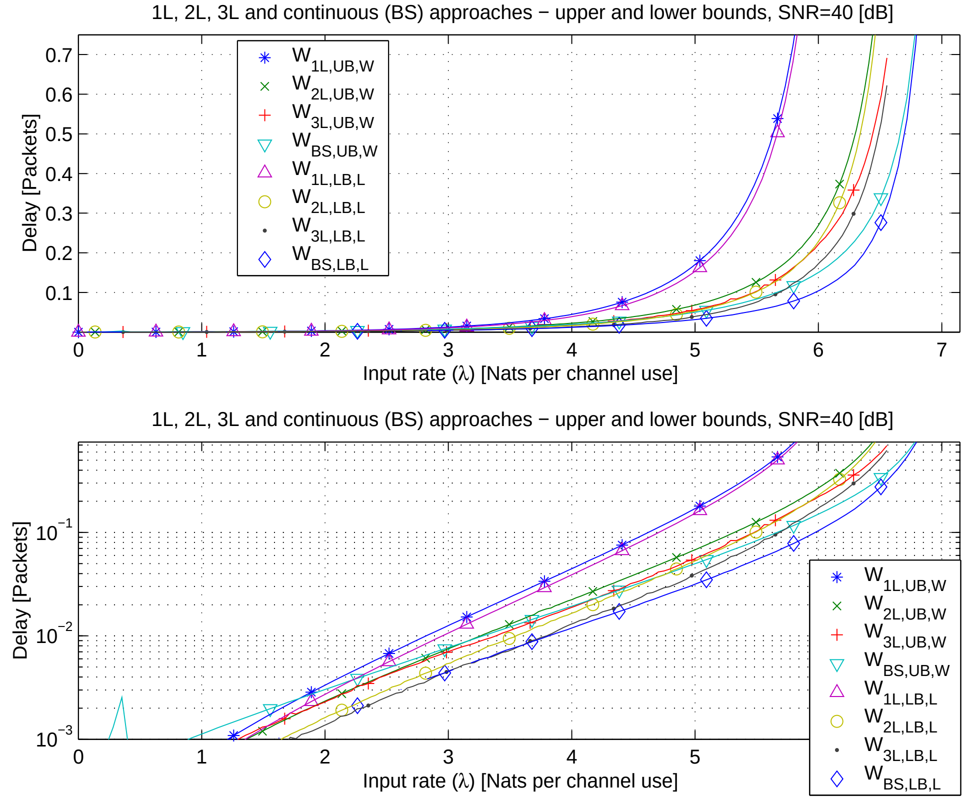

Fig. 10. Average delay for outage approach, 2-level and 3-level coding and continuous layering (BS), for SNR=40dB. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

Taking the laplace transform of the PDF of (A.1) is required as an initial step of deriving the bounds,

$$
\begin{aligned}
L_W(s) &= \int_0^\infty e^{-sw} dF_W(w) \\
&\overset{(1)}{=} F_W(0) + p_1 \int_0^\infty e^{-sw} dF_W(w-\lambda+R_1+R_2) + p_2 \int_{\lambda-R_1}^\infty e^{-sw} dF_W(w-\lambda+R_1) \\
&\quad +\overline{p} \int_\lambda^\infty e^{-sw} dF_W(w-\lambda) \\
&\overset{(2)}{=} F_W(0) + p_1 \int_{R_1+R_2-\lambda}^\infty e^{-s(w+\lambda-R_1-R_2)} dF_W(w) + p_2 \int_0^\infty e^{-s(w+\lambda-R_1)} dF_W(w) \\
&\quad +\overline{p} \int_0^\infty e^{-s(w-\lambda)} dF_W(w) \qquad\qquad \text{(A.2)} \\
&\overset{(3)}{=} F_W(0) + [p_1 e^{-s(\lambda-R_1-R_2)} + p_2 e^{-s(\lambda-R_1)} + \overline{p} e^{-s\lambda}] \int_0^\infty e^{-sw} dF_W(w) \\
&\quad -p_1 \int_0^{R_1+R_2-\lambda} e^{-s(w+\lambda-R_1-R_2)} dF_W(w) \\
&\overset{(4)}{=} F_W(0) + [p_1 e^{-s(\lambda-R_1-R_2)} + p_2 e^{-s(\lambda-R_1)} + \overline{p} e^{-s\lambda}] L_W(s) \\
&\quad -p_1 \int_0^{R_1+R_2-\lambda} e^{-s(w+\lambda-R_1-R_2)} dF_W(w)
\end{aligned}
$$

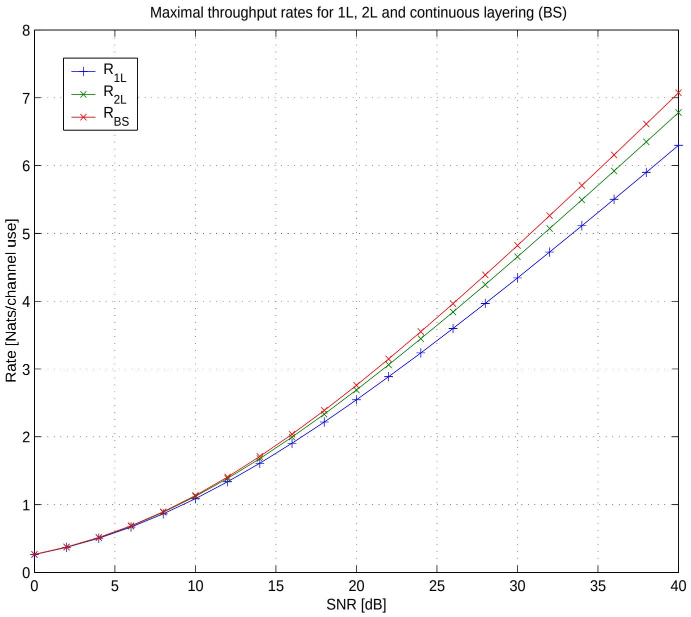

Fig. 11. Maximal throughput vs. SNR, for outage approach, 2-level coding and continuous layering (BS).

where the PDF of $dF_W(w)$ is denoted $L_W(s)$. (1) substituting the right-hand side of (A.1) into the Laplace transform definition. (2) change of integral variables. (3) Taking the integral as a common factor. (4) Substituting back the integral definition with the Laplace transform. The last step suggests a new expression for the Laplace transform of $dF_W(w)$. That is $L_W(s)$ can now be expressed by

$$\begin{aligned} L_W(s) &= \frac{F_W(0) - p_1 \int_0^{R_1+R_2-\lambda} e^{-s(w+\lambda-R_1-R_2)} dF_W(w)}{1 - [p_1 e^{-s(\lambda-R_1-R_2)} + p_2 e^{-s(\lambda-R_1)} + \overline{p} e^{-s\lambda}]} \\ &=^{(1)} \frac{p_1 \int_0^{R_1+R_2-\lambda} (1 - e^{-s(w+\lambda-R_1-R_2)}) dF_W(w)}{1 - [p_1 e^{-s(\lambda-R_1-R_2)} + p_2 e^{-s(\lambda-R_1)} + \overline{p} e^{-s\lambda}]} \\ &=^{(2)} \frac{p_1 \int_0^{R_1+R_2-\lambda} (e^{-s(R_1+R_2-\lambda)} - e^{-sw}) dF_W(w)}{e^{-s(R_1+R_2-\lambda)} - [p_1 + p_2 e^{-sR_2} + \overline{p} e^{-s(R_1+R_2)}]} \\ &\triangleq \frac{L_Y(s)}{L_X(s)}, \end{aligned} \tag{A.3}$$

where the first equation is a direct substitution of $L_W(s)$ from (A.2). (1) involves replacement of $F_W(0)$ with its equivalent from eq. (A.1), that is $F_W(0) = p_1 F_W(w - (\lambda - R_1 - R_2))$. (2) multiplication of numerator and denominator by a common factor.

Generally, the first moment of W is given by

$$E[W] = \lim_{s \to 0} -\frac{dL_W(s)}{ds}, \tag{A.4}$$

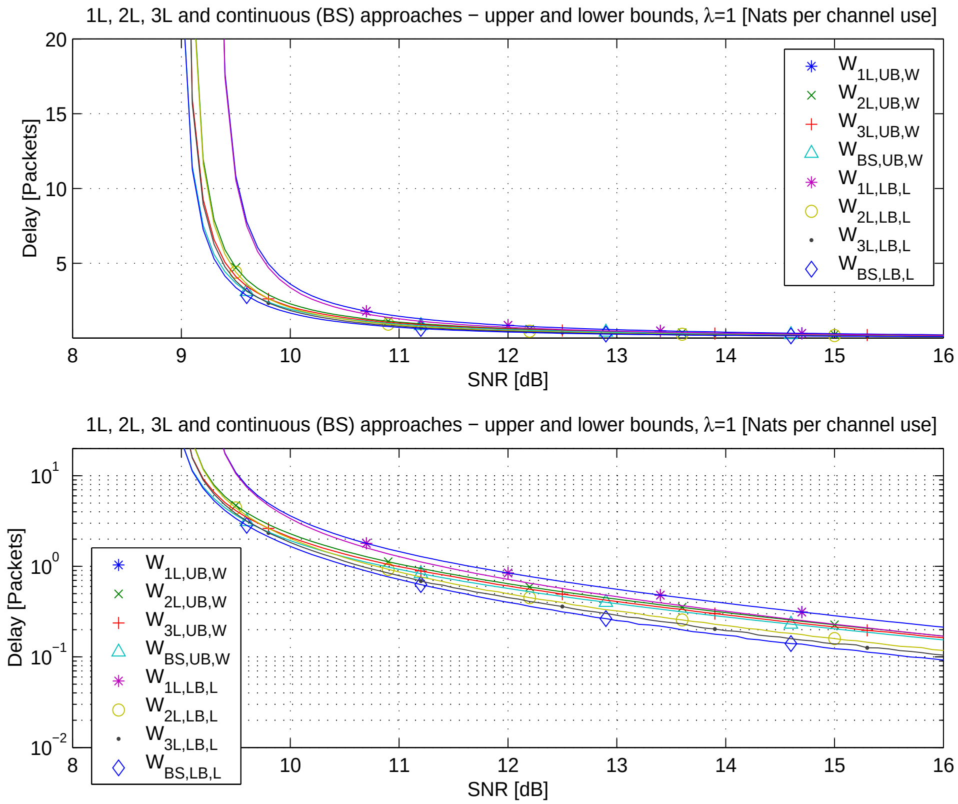

Fig. 12. Average delay vs. SNR, for outage approach, 2-level, 3-level coding and continuous layering (BS), for $\lambda = 1$ [Nats/channel use]. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

however in (A.3) we can see that for $s = 0$ both $L_Y(s)|_{s=0} = 0$ and $L_X(s)|_{s=0} = 0$. Therefore, L'Hospital rule for a fraction, for which numerator and denominator tend to zero, should be used. This requires assuming that $L_Y(s)$ and $L_X(s)$ have second order derivatives. Applying

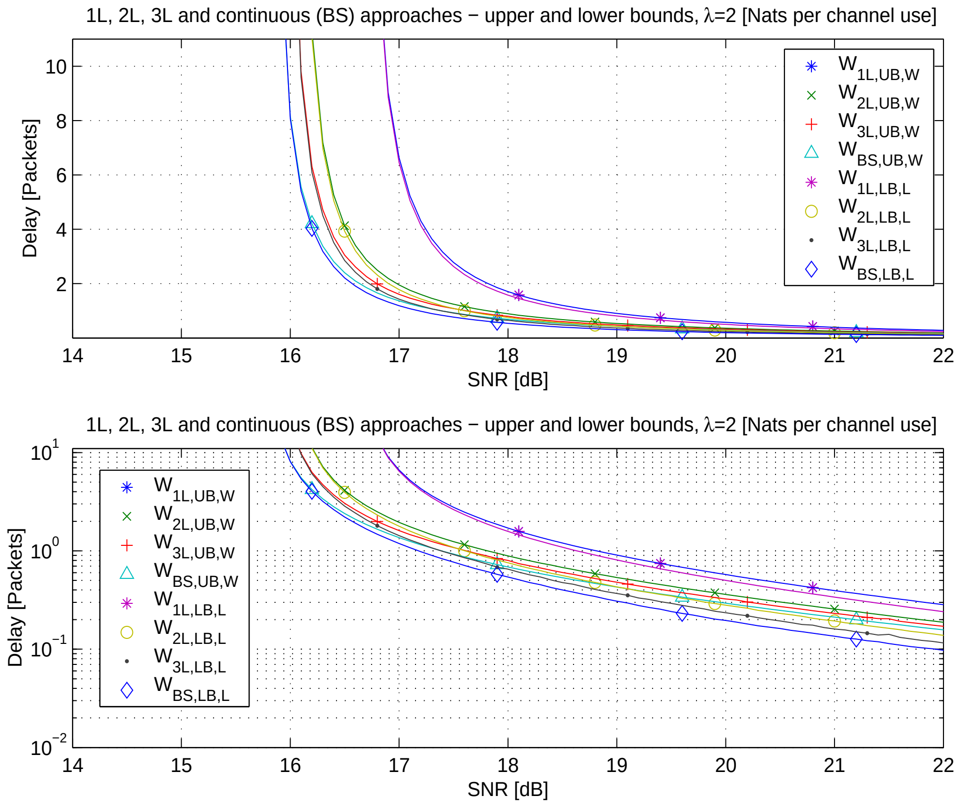

Fig. 13. Average delay vs. SNR, for outage approach, 2-level, 3-level coding and continuous layering (BS), for $\lambda = 2$ [Nats/channel use]. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

L'Hospital rule on (A.3),

$$
\begin{aligned}
E[W] &= \lim_{s\to 0} -\frac{dL_W(s)}{ds} \\
&\overset{(1)}{=} \lim_{s\to 0} -\frac{L'_Y(s)L_X(s)-L'_X(s)L_Y(s)}{L^2_X(s)} \\
&\overset{(2)}{=} \lim_{s\to 0} \frac{L''_X(s)L_Y(s)+L'_X(s)L'_Y(s)-L''_Y(s)L_X(s)-L'_X(s)L'_Y(s)}{2L_X(s)L'_X(s)} \\
&\overset{(3)}{=} \lim_{s\to 0} \frac{L'''_X(s)L_Y(s)+L''_X(s)L'_Y(s)-L'''_Y(s)L_X(s)-L'_X(s)L''_Y(s)}{2[(L'_X(s))^2+L_X(s)L''_X(s)]} \\
&\overset{(4)}{=} \lim_{s\to 0} \frac{L''_X(s)L'_Y(s)-L'_X(s)L''_Y(s)}{2(L'_X(s))^2} \\
&\overset{(5)}{=} \lim_{s\to 0} \frac{L''_X(s)-L''_Y(s)}{2L'_X(s)}
\end{aligned}
\tag{A.5}
$$

where $L'_X(s)$ represents the first order derivative of $L_X(s)$ w.r.t. s. (1) a derivative of a fraction,

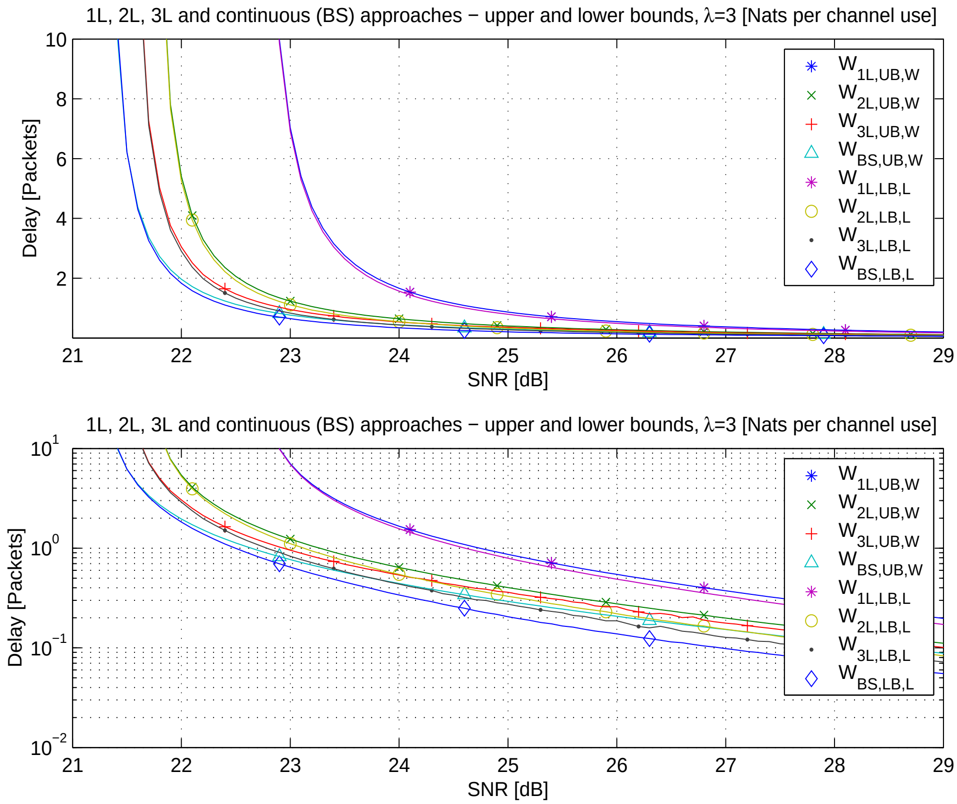

Fig. 14. Average delay vs. SNR, for outage approach, 2-level, 3-level coding and continuous layering (BS), for $\lambda = 3$ [Nats/channel use]. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

by definition. (2) applying L'Hospital rule, derivation of numerator and denominator separately. (3) uses $L_Y(s)|_{s=0} = 0$ and $L_X(s)|_{s=0} = 0$, and performing another L'Hospital derivation. (4) uses again $L_Y(s)|_{s=0} = 0$ and $L_X(s)|_{s=0} = 0$. (5) comes from the fact that $\lim_{s\to 0} L_W(s) = 0$, which suggests

$$\lim_{s\to 0}\frac{L'_Y(s)}{L'_X(s)} = 1 \tag{A.6}$$

In this stage the first order derivatives of $L_Y(s)$ and $L_X(s)$ are computed.

$$\begin{aligned}
&\lim_{s\to 0} L'_Y(s) = \lim_{s\to 0} p_1 \int_0^{R_1+R_2-\lambda}[we^{-sw} - (R_1+R_2-\lambda)e^{-s(R_1+R_2-\lambda)}]dF_W(w)\\
&= p_1 \int_0^{R_1+R_2-\lambda}(w - R_1 - R_2 + \lambda)dF_W(w)\\
&\overset{(1)}{=} \; -p_1 \int_0^{R_1+R_2-\lambda} F_W(w)dw
\end{aligned} \tag{A.7}$$

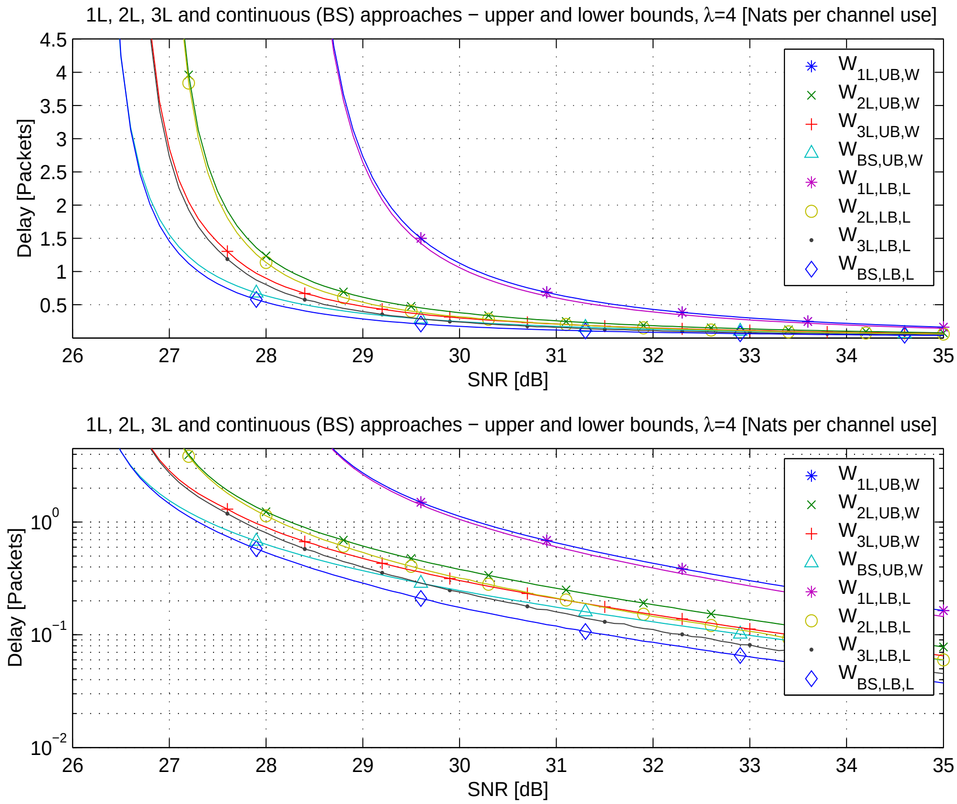

Fig. 15. Average delay vs. SNR, for outage approach, 2-level, 3-level coding and continuous layering (BS), for $\lambda = 4$ [Nats/channel use]. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

where (1) is a result of solving the integral in parts.

$$\begin{aligned}&\lim_{s\to 0} L'_X(s) = \lim_{s\to 0} -(R_1+R_2-\lambda)e^{-s(R_1+R_2-\lambda)} + p_2R_2e^{-sR_2} + \overline{p}(R_1+R_2)e^{-s(R_1+R_2)}\\&= \lambda - p_1(R_1+R_2) - p_2R_1.\end{aligned} \tag{A.8}$$

Taking the recent two equations (A.7) and (A.8), substituting the derivatives in (A.6) results in a useful equality

$$p_1\int_0^{R_1+R_2-\lambda} F_W(w)dw = p_1(R_1+R_2) + p_2R_1 - \lambda. \tag{A.9}$$

Now the second order derivatives of $L_Y(s)$ and $L_X(s)$ are required. From (A.8), $L''_X(s)|_{s=0}$ is directly derived

$$L''_X(s)|_{s=0} = (R_1+R_2-\lambda)^2 - p_2R_2^2 - \overline{p}(R_1+R_2)^2, \tag{A.10}$$

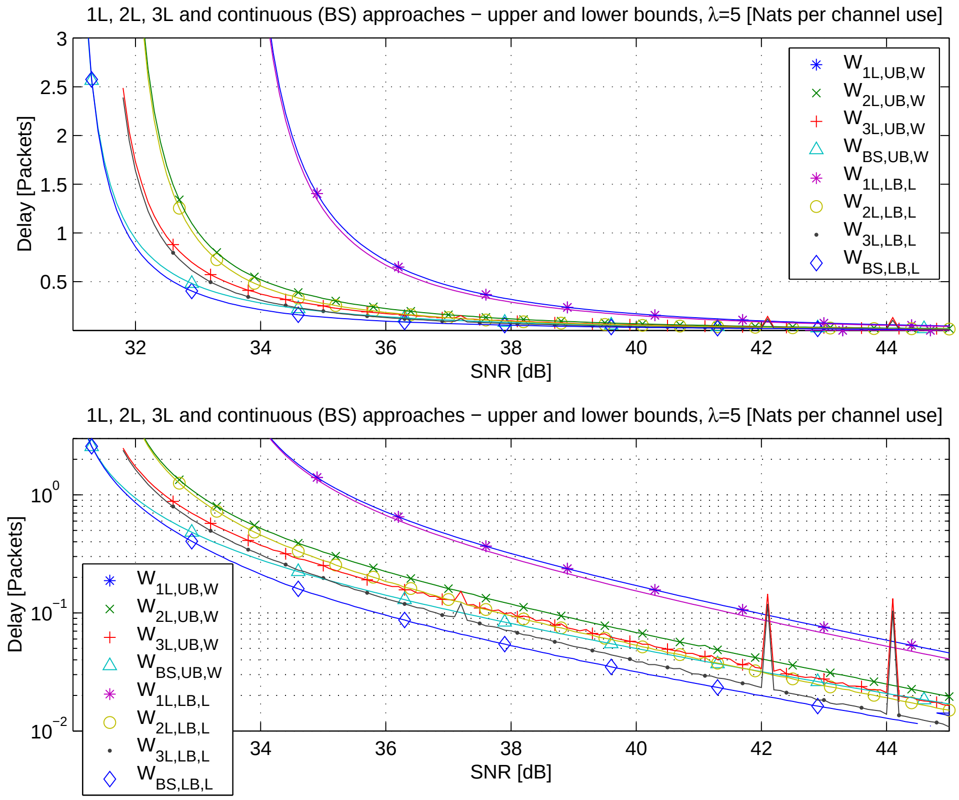

Fig. 16. Average delay vs. SNR, for outage approach, 2-level, 3-level coding and continuous layering (BS), for $\lambda = 5$ [Nats/channel use]. The bounds $W_{nL,LB,L}$ denote the n-level coding **lower** bounds specified in (42). The bounds $W_{nL,UB,W}$ denote the n-level coding **upper** bounds specified in its general form in (13).

calculating $L''_Y(s)|_{s=0}$ will allow also to compute the bounds.

$$\begin{aligned} L''_Y(s)|_{s=0} &= p_1 \int_0^{R_1+R_2-\lambda}[-w^2 e^{-sw} + (R_1+R_2-\lambda)^2 e^{-s(R_1+R_2-\lambda)}]dF_W(w)|_{s=0} \\ &= p_1 \int_0^{R_1+R_2-\lambda}[-w^2 + (R_1+R_2-\lambda)^2]dF_W(w) \\ &= p_1 \int_0^{R_1+R_2-\lambda}(R_1+R_2-\lambda-w)(R_1+R_2-\lambda+w)dF_W(w). \end{aligned} \tag{A.11}$$

The last equation in (A.11) can be upper bounded by replacing $(R_1+R_2-\lambda+w)$ with $2(R_1+R_2-\lambda)$, that is

$$\begin{aligned} L''_Y(s)|_{s=0} &\leq 2(R_1+R_2-\lambda)p_1 \int_0^{R_1+R_2-\lambda}(R_1+R_2-\lambda-w)dF_W(w) \\ &\overset{(1)}{=} 2(R_1+R_2-\lambda)(p_1(R_1+R_2)+p_2R_1-\lambda) \end{aligned} \tag{A.12}$$

where we have used (A.9) in step (1) to obtain an explicit expression for the upper bound on $L''_Y(s)|_{s=0}$. Substituting this bound together with (A.8) and (A.10) into (A.5) we reach

$$EW \leq \frac{2(R_1+R_2-\lambda)(p_1(R_1+R_2)+p_2R_1-\lambda) - (R_1+R_2-\lambda)^2 + p_2R_2^2 + \overline{p}(R_1+R_2)^2}{2(p_1(R_1+R_2)+p_2R_1-\lambda)} \tag{A.13}$$

which after some algebra reduces to (29).

Similarly a lower bound of $L''_Y(s)|_{s=0}$ is obtained by replacing $(R_1 + R_2 - \lambda + w)$ with $R_1 + R_2 - \lambda$ in (A.11),

$$\begin{array}{l} L''_Y(s)|_{s=0} \geq p_1(R_1 + R_2 - \lambda) \int_0^{R_1+R_2-\lambda} (R_1 + R_2 - \lambda - w) dF_W(w) \\ (R_1 + R_2 - \lambda)(p_1(R_1 + R_2) + p_2 R_1 - \lambda). \end{array} \tag{A.14}$$

where we have used (A.9) in step (1) again to obtain an explicit expression for the lower bound on $L''_Y(s)|_{s=0}$. Substituting this bound together with (A.8) and (A.10) into (A.5) we reach

$$EW \geq \frac{(R_1 + R_2 - \lambda)(p_1(R_1 + R_2) + p_2 R_1 - \lambda) - (R_1 + R_2 - \lambda)^2 + p_2 R_2^2 + \overline{p}(R_1 + R_2)^2}{2(p_1(R_1 + R_2) + p_2 R_1 - \lambda)} \tag{A.15}$$

which after some algebra reduces to (28). It takes only normalization by λ to reach (30) and (31) from (28) and (29) respectively. This concludes **part A** of the proof, for the case that $R_1 \leq \lambda$.

B) $R_1 > \lambda$**:** The queue size CDF follows from (26) and (27), differs from (A.1),

$$F_W(w) = \begin{cases} 0 & w < 0 \\ p_1 F_W(w - (\lambda - R_1 - R_2)) + p_2 F_W(w - (\lambda - R_1)) & 0 \leq w \leq \lambda \\ p_1 F_W(w - (\lambda - R_1 - R_2)) + p_2 F_W(w - (\lambda - R_1)) + \overline{p} F_W(w - \lambda) & \lambda \leq w \end{cases} \tag{A.16}$$

Taking the laplace transform of the PDF in both of (A.16) is required as an initial step of deriving the bounds,

$$\begin{array}{rcl} L_W(s) & = & \int\limits_0^\infty e^{-sw} dF_W(w) \\ & \overset{(1)}{=} & F_W(0) + p_1 \int\limits_0^\infty e^{-sw} dF_W(w - \lambda + R_1 + R_2) + p_2 \int\limits_0^\infty e^{-sw} dF_W(w - \lambda + R_1) \\ & & + \overline{p} \int\limits_\lambda^\infty e^{-sw} dF_W(w - \lambda) \\ & \overset{(2)}{=} & F_W(0) + p_1 \int\limits_{R_1+R_2-\lambda}^\infty e^{-s(w+\lambda-R_1-R_2)} dF_W(w) + p_2 \int\limits_{R_1-\lambda}^\infty e^{-s(w+\lambda-R_1)} dF_W(w) \\ & & + \overline{p} \int\limits_0^\infty e^{-s(w-\lambda)} dF_W(w) \\ & \overset{(3)}{=} & F_W(0) + [p_1 e^{-s(\lambda-R_1-R_2)} + p_2 e^{-s(\lambda-R_1)} + \overline{p} e^{-s\lambda}] L_W(s) \\ & & - p_1 \int\limits_0^{R_1+R_2-\lambda} e^{-s(w+\lambda-R_1-R_2)} dF_W(w) - p_2 \int\limits_0^{R_1-\lambda} e^{-s(w+\lambda-R_1)} dF_W(w) \end{array} \tag{A.17}$$

where (1) is substituting the right-hand side of (A.16) into the Laplace transform definition. (2) change of integral variables. (3) substituting back the integral definition with the Laplace transform and taking a common factor. The last step suggests a new expression for the Laplace transform of $dF_W(w)$, similarly to (A.3),

$$\begin{aligned} L_W(s) = & \frac{p_1 \int_0^{R_1+R_2-\lambda}(e^{-s(R_1+R_2-\lambda)} - e^{-sw})dF_W(w) + p_2 \int_0^{R_1-\lambda}(e^{-s(R_1+R_2-\lambda)} - e^{-s(w+R_2)})dF_W(w)}{e^{-s(R_1+R_2-\lambda)} - [p_1 + p_2 e^{-sR_2} + \overline{p} e^{-s(R_1+R_2)}]} \\ \triangleq & \frac{L_Z(s)}{L_X(s)}, \end{aligned} \quad \text{(A.18)}$$

where we have used the equality $F_W(0) = p_1 F_W(w - (\lambda - R_1 - R_2)) + p_2 F_W(w - (\lambda - R_1))$ from (A.16). As may be noticed here the denominator is exactly the same $L_X(s)$ of () in part A of the proof, which means that the denominator of $L_W(s)$ is independent on the value of R_1 relative to λ. We use here (A.18) to bound $E[W]$. Here again, both $L_Z(s)|_{s=0} = 0$ and $L_X(s)|_{s=0} = 0$. Therefore we will use L'Hospital rule and the result of (A.5). Expressions of $L'_X(s)$ and $L''_X(s)$ are already stated in (A.8) and (A.10) respectively.

$$\begin{aligned} L'_Z(s)|_{s=0} &= \{p_1 \int_0^{R_1+R_2-\lambda}[we^{-sw} - (R_1+R_2-\lambda)e^{-s(R_1+R_2-\lambda)}]dF_W(w) \\ &\quad + p_2 \int_0^{R_1-\lambda}((w+R_2)e^{-s(w+R_2)} - (R_1+R_2-\lambda)e^{-s(R_1+R_2-\lambda)})dF_W(w)\} \ |_{s=0} \\ &= p_1 \int_0^{R_1+R_2-\lambda}(w - R_1 - R_2 + \lambda)dF_W(w) + p_2 \int_0^{R_1-\lambda}(w - R_1 + \lambda)dF_W(w) \\ &\overset{(1)}{=} -p_1 \int_0^{R_1+R_2-\lambda} F_W(w)dw - p_2 \int_0^{R_1-\lambda} F_W(w)dw \end{aligned} \quad \text{(A.19)}$$

Substituting (A.8) and (A.19) into (A.6) we reach an equality similar to (A.9),

$$p_1 \int_0^{R_1+R_2-\lambda} F_W(w)dw + p_2 \int_0^{R_1-\lambda} F_W(w)dw = p_1(R_1+R_2) + p_2 R_1 - \lambda. \quad \text{(A.20)}$$

Now the second order derivative of $L_Z(s)$ is required.

$$\begin{aligned} L''_Z(s)|_{s=0} &= \{p_1 \int_0^{R_1+R_2-\lambda}[-w^2 e^{-sw} + (R_1+R_2-\lambda)^2 e^{-s(R_1+R_2-\lambda)}]dF_W(w) \\ &\quad + p_2 \int_0^{R_1-\lambda}(-(w+R_2)^2 e^{-s(w+R_2)} + (R_1+R_2-\lambda)^2 e^{-s(R_1+R_2-\lambda)})dF_W(w)\} \ |_{s=0} \\ &= p_1 \int_0^{R_1+R_2-\lambda}(R_1+R_2-\lambda-w)(R_1+R_2-\lambda+w)dF_W(w) \\ &\quad + p_2 \int_0^{R_1-\lambda}(w + R_1 + 2R_2 - \lambda)(R_1 - \lambda - w)dF_W(w). \end{aligned} \quad \text{(A.21)}$$

The last expression of $L''_Z(s)|_{s=0}$ in (A.21) can be upper bounded by replacing w with its maximal values in both integrals. That is substitute $(R_1 + R_2 - \lambda + w)$ by $2(R_1 + R_2 - \lambda)$ and also $(R_1 + 2R_2 - \lambda + w)$ is substituted by $2(R_1 + R_2 - \lambda)$, thus

$$\begin{aligned} L''_Z(s)|_{s=0} &\leq 2p_1(R_1+R_2-\lambda)\int_0^{R_1+R_2-\lambda}(R_1+R_2-\lambda-w)dF_W(w) \\ &\quad + 2p_2(R_1+R_2-\lambda)\int_0^{R_1-\lambda}(R_1-\lambda-w)dF_W(w). \\ &\overset{(1)}{=} 2(R_1+R_2-\lambda)(p_1(R_1+R_2) + p_2 R_1 - \lambda) \end{aligned} \quad \text{(A.22)}$$

where we have used (A.20) in step (1) to obtain an explicit expression for the upper bound on $L''_Z(s)|_{s=0}$. Substituting this bound together with (A.8) and (A.10) into (A.3) we reach the same upper bound as in (A.13),

$$EW \leq \frac{2(R_1+R_2-\lambda)(p_1(R_1+R_2)+p_2R_1-\lambda)-(R_1+R_2-\lambda)^2+p_2R_2^2+\overline{p}(R_1+R_2)^2}{2(p_1(R_1+R_2)+p_2R_1-\lambda)} \tag{A.23}$$

which after some algebra reduces to (29). Similarly a lower bound of $L''_Z(s)|_{s=0}$ is obtained by substituting $(R_1+R_2-\lambda+w)$ by $(R_1+R_2-\lambda)$ and by replacing $(R_1+2R_2-\lambda+w)$ also by $(R_1+R_2-\lambda)$ in the second integral of (A.21),

$$\begin{aligned} L''_Z(s)|_{s=0} &\geq p_1(R_1+R_2-\lambda)\int_0^{R_1+R_2-\lambda}(R_1+R_2-\lambda-w)dF_W(w) \\ &\quad +p_2(R_1+R_2-\lambda)\int_0^{R_1-\lambda}(R_1-\lambda-w)dF_W(w). \\ &\overset{(1)}{=} (R_1+R_2-\lambda)(p_1(R_1+R_2)+p_2R_1-\lambda) \end{aligned} \tag{A.24}$$

where we have used (A.20) in step (1) again to obtain an explicit expression for the lower bound on $L''_Z(s)|_{s=0}$. Substituting this bound together with (A.8) and (A.10) into (A.3) we reach the same lower bound as in (A.15),

$$EW \geq \frac{(R_1+R_2-\lambda)(p_1(R_1+R_2)+p_2R_1-\lambda)-(R_1+R_2-\lambda)^2+p_2R_2^2+\overline{p}(R_1+R_2)^2}{2(p_1(R_1+R_2)+p_2R_1-\lambda)} \tag{A.25}$$

which after some algebra reduces to (28). It takes only normalization by λ to reach (30) and (31) from (28) and (29) respectively.

This shows that in both parts for either $R_1 > \lambda$ and $R_1 \leq \lambda$ the same upper and lower bounds on the expected waiting time are valid, although the CDF $F_W(w)$ is different in these two cases. ■

APPENDIX B

PROOF OF THEOREM 6.1

Proof: The main steps of the proof resemble the two level layering. Only here we assume that there is some k, $1 \leq k \leq K$ for which

$$\sum_{i=1}^{k} R_i \leq \lambda \leq \sum_{i=1}^{k+1} R_i \tag{B.1}$$

The queue size CDF follows from (26) and (39),

$$F_W(w) = \begin{cases} 0 & w < 0 \\ \sum\limits_{i=1}^{K-k} p_i F_W(w - (\lambda - \sum_{j=1}^{K-i+1} R_j)) & 0 \leq w < \lambda - \sum_{i=1}^{k} R_i \\ \sum\limits_{i=1}^{K-k+1} p_i F_W(w - (\lambda - \sum_{j=1}^{K-i+1} R_j)) & \sum_{i=1}^{k} R_i \leq w < \lambda - \sum_{i=1}^{k-1} R_i \\ \vdots & \vdots \\ \sum\limits_{i=1}^{K} p_i F_W(w - (\lambda - \sum_{j=1}^{K-i+1} R_j)) & \sum_{i=1}^{k} R_i \leq w < \lambda \\ \sum\limits_{i=1}^{K} p_i F_W(w - (\lambda - \sum_{j=1}^{K-i+1} R_j)) + \overline{p} F_W(w - \lambda) & \lambda \leq w \end{cases} \tag{B.2}$$

For compactness of presentation $\sum_{j=1}^{V} R_j$ will be denoted $\Re_V \triangleq \sum_{j=1}^{V} R_j$. Taking the laplace transform of the PDF in both of (B.2) is required as an initial step of deriving the bounds,

$$\begin{aligned} L_W(s) &= \int\limits_0^\infty e^{-sw} dF_W(w) \\ &\overset{(1)}{=} F_W(0) + \int\limits_0^\infty e^{-sw} \sum_{i=1}^{K-k} p_i dF_W(w - \lambda + \Re_{K-i+1}) + p_{K-k+1} \int\limits_{\lambda-\Re_k}^\infty e^{-sw} dF_W(w - \lambda + \Re_k) + \cdots \\ &\quad p_K \int\limits_{\lambda - R_1}^\infty e^{-sw} dF_W(w - \lambda + R_1) + \overline{p} \int\limits_\lambda^\infty e^{-sw} dF_W(w - \lambda) \\ &\overset{(2)}{=} F_W(0) + [\sum_{i=1}^{K} p_i e^{s(\Re_{K-i+1} - \lambda)} + \overline{p} e^{-s\lambda}] L_W(s) \\ &\quad - \sum_{i=1}^{K-k} p_i \int\limits_0^{\Re_{K-i+1}-\lambda} e^{-s(w+\lambda-\Re_{K-i+1})} dF_W(w) \end{aligned} \tag{B.3}$$

where the PDF of $dF_W(w)$ is denoted $L_W(s)$. (1) substituting the right-hand side of (B.2) into the Laplace transform definition. (2) change of integral variables and takes the integral as a common factor. The last step suggests a new expression for the Laplace transform of $dF_W(w)$. That is $L_W(s)$ can now be expressed by

$$\begin{aligned} L_W(s) &= \frac{F_W(0) - \sum\limits_{i=1}^{K-k} p_i \int\limits_0^{\Re_{K-i+1}-\lambda} e^{-s(w+\lambda-\Re_{K-i+1})} dF_W(w)}{1 - [\sum\limits_{i=1}^{K} p_i e^{s(\Re_{K-i+1}-\lambda)} + \overline{p} e^{-s\lambda}]} \\ &\overset{(1)}{=} \frac{\sum\limits_{i=1}^{K-k} p_i \int\limits_0^{\Re_{K-i+1}-\lambda} (e^{-s(\Re_K - \lambda)} - e^{-s(w + \Re_K - \Re_{K-i+1})}) dF_W(w)}{e^{-s(\Re_K - \lambda)} - [\sum\limits_{i=1}^{K} p_i e^{s(\Re_K - \Re_{K-i+1})} + \overline{p} e^{-s\Re_K}]} \\ &\triangleq \frac{L_Y(s)}{L_X(s)} \end{aligned} \tag{B.4}$$

where the first equation is a direct substitution of $L_W(s)$ from (B.3). (1) involves replacement of $F_W(0)$ with its equivalent from eq. (B.2), that is $F_W(0) = \sum_{i=1}^{K-k} p_i F_W(-\lambda + \Re_{K-i+1})$. (2) multiplication of numerator and denominator by a common factor.

In general the first moment of W is specified in (A.4), however in (B.4) it can be noticed that for $s = 0$ both $L_Y(s)|_{s=0} = 0$ and $L_X(s)|_{s=0} = 0$, like in the two level coding case. Therefore L'Hospital rule is used here as well. This requires assuming that $L_Y(s)$ and $L_X(s)$ have second order derivatives. The result of $E[W]$ specified by the first and second order derivatives in (A.5), shall be used in the following.

In this stage the first order derivatives of $L_Y(s)$ and $L_X(s)$ are given by

$$\begin{aligned} L'_Y(s)|_{s=0} &= \sum_{i=1}^{K-k} p_i \int\limits_0^{\Re_{K-i+1}-\lambda} [(w + \Re_K - \Re_{K-i+1})e^{-s(w+\Re_K-\Re_{K-i+1})} - (\Re_K - \lambda)e^{-s(\Re_K-\lambda)}]dF_W(w)|_{s=0} \\ &= \sum_{i=1}^{K-k} p_i \int\limits_0^{\Re_{K-i+1}-\lambda} (w - \Re_{K-i+1} + \lambda))dF_W(w) \\ &= -\sum_{i=1}^{K-k} p_i \int\limits_0^{\Re_{K-i+1}-\lambda} F_W(w)dw \end{aligned} \tag*{(B.5}$$

$$\begin{aligned} L'_X(s)|_{s=0} &= -(\Re_K - \lambda)e^{-s(\Re_K-\lambda)} + \sum_{i=1}^{K} p_i(\Re_K - \Re_{K-i+1})e^{-s(\Re_K-\Re_{K-i+1})} + \Re_K\overline{p}e^{-s\Re_K}|_{s=0} \\ &= -\Re_K + \lambda + \sum_{i=1}^{K} p_i(\Re_K - \Re_{K-i+1}) + \Re_K\overline{p} \\ &= \lambda - \sum_{i=1}^{K} p_i\Re_{K-i+1} \end{aligned} \tag{B.6}$$

Taking the recent two equations (B.5) and (B.6), substituting the derivatives in (A.6) results in a useful equality

$$\sum_{i=1}^{K-k} p_i \int\limits_0^{\Re_{K-i+1}-\lambda} F_W(w)dw = \sum_{i=1}^{K} p_i\Re_{K-i+1} - \lambda. \tag{B.7}$$

Now the second order derivatives of $L_Y(s)$ and $L_X(s)$ are required. From (B.6), $L''_X(s)|_{s=0}$ is directly derived

$$L''_X(s)|_{s=0} = (\Re_K - \lambda)^2 - \sum_{i=1}^{K} p_i(\Re_K - \Re_{K-i+1})^2 - \overline{p}\Re_K^2 \tag{B.8}$$

calculating $L''_Y(s)|_{s=0}$ will allow also to compute the bounds.

$$\begin{aligned} L''_Y(s)|_{s=0} &= \sum_{i=1}^{K-k} p_i \int\limits_0^{\Re_{K-i+1}-\lambda} [-(w + \Re_K - \Re_{K-i+1})^2 + (\Re_K - \lambda)^2]dF_W(w) \\ &= \sum_{i=1}^{K-k} p_i \int\limits_0^{\Re_{K-i+1}-\lambda} (w + 2\Re_K - \Re_{K-i+1} - \lambda)(\Re_{K-i+1} - \lambda - w)dF_W(w) \end{aligned} \tag{B.9}$$

The last equation in (B.9) can be upper bounded by replacing w of $(w + 2\Re_K + \Re_{K-i+1} - \lambda)$ by $\Re_{K-i+1} - \lambda$,

$$\begin{aligned} L''_Y(s)|_{s=0} &\leq 2(\Re_K - \lambda) \sum_{i=1}^{K-k} p_i \int_0^{\Re_{K-i+1}-\lambda} (\Re_{K-i+1} - \lambda - w) dF_W(w) \\ &= 2(\Re_K - \lambda)(\sum_{i=1}^{K} p_i \Re_{K-i+1} - \lambda) \end{aligned} \tag{B.10}$$

where we have used (B.7) in the final step to obtain an explicit expression for the upper bound of $L''_Y(s)|_{s=0}$. Substituting this bound together with (B.6) and (B.8) into (A.5) we reach

$$EW \leq \frac{2(\Re_K - \lambda)(\sum_{i=1}^{K} p_i \Re_{K-i+1} - \lambda) - (\Re_K - \lambda)^2 + \sum_{i=1}^{K} p_i (\Re_K - \Re_{K-i+1})^2 + \overline{p}\Re_K^2}{2(\sum_{i=1}^{K} p_i \Re_{K-i+1} - \lambda)} \tag{B.11}$$

which is exactly the desired upper bound in (41). Similarly a lower bound of $L''_Y(s)|_{s=0}$ is obtained from (B.9) by replacing w of $(w + 2\Re_K + \Re_{K-i+1} - \lambda)$ with $\Re_{K-i+1} - \Re_K$,

$$\begin{aligned} L''_Y(s)|_{s=0} &\geq (\Re_K - \lambda) \sum_{i=1}^{K-k} p_i \int_0^{\Re_{K-i+1}-\lambda} (\Re_{K-i+1} - \lambda - w) dF_W(w) \\ &= (\Re_K - \lambda)(\sum_{i=1}^{K} p_i \Re_{K-i+1} - \lambda) \end{aligned} \tag{B.12}$$

where we have used (B.7) in the last step again to obtain an explicit expression for the lower bound on $L''_Y(s)|_{s=0}$. Substituting this bound together with (B.6) and (B.8) into (A.5) we reach

$$EW \geq \frac{(\Re_K - \lambda)(\sum_{i=1}^{K} p_i \Re_{K-i+1} - \lambda) - (\Re_K - \lambda)^2 + \sum_{i=1}^{K} p_i (\Re_K - \Re_{K-i+1})^2 + \overline{p}\Re_K^2}{2(\sum_{i=1}^{K} p_i \Re_{K-i+1} - \lambda)} \tag{B.13}$$

which is exactly the desired lower bound (40). It takes only normalization by λ to reach (42) and (43) from (40) and (41) respectively. ■

REFERENCES

[1] E. Telatar and R. G. Gallager, “Combining netork thoery and information theory for multi-access,” *IEEE Selected areas in communications*, vol. 13, pp. 963–969, August 1995.

[2] A. Ephremides and B. Hajek, “Information theory and communication networks: An unconsummated union,” *IEEE Trans. on Inform. Theory*, vol. 44, no. 3, pp. 2416–2434, July 1998.

[3] R. G. Gallager, “A perspective on multiaccess channels,” *IEEE Trans. on Inform. Theory*, vol. 31, no. 2, pp. 124–142, March 1985.

[4] T. Coleman and M. Medard, “The impact of user information on power-delay tradeoffs between in bursty packetized systems,” *IEEE Int. Symp. Inform. Theory (ISIT’03), Yokohoma, Japan*, p. 811, June 29 - July 4 2003.

[5] V. Anantharam and S. Verdu, “Bits through queues,” *IEEE Trans. on Info, Theory*, vol. 42, no. 1, pp. 4–18, January 1996.

[6] R. Berry, “Power and delay trade-offs in fading channels,” *PhD Thesis dissertation, MIT*, June 2000.

[7] R. A. Berry and R. G. Gallager, “Communication over fading channels with delay constraints,” *IEEE Trans. on Inform. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[8] M. Goyal, A. Kumar, and V. Sharma, “Power constrained and delay optimal policies for scheduling transmission over a fading channel,” *The 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), San-Francisco*, April 1 - 3 2003.

[9] Y.-C. Liang and R. Zhang, “Transmit optimization for mimo channels with mixed delay-constrained and no-delay-constrained services,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 17 - 21 2004.

[10] I. Bettesh and S. Shamai (Shitz), “Optimal power and rate control for minimal average delay: the single-user case,” *submitted to IEEE Trans. on Inform. Theory*, December 2001.

[11] I. Bettesh, “Information and network theory aspects of communication systems in fading enviornment,” *Thesis dissertation*, September 2002.

[12] D. P. Bertsekas, *Dynamic Programming and Optimal Control Vol 1,2*. Belmont, Massachusetts: Athena Scientific, 1995.

[13] B. E. Collins and R. L. Cruz, “Transmission policies for time varying channels with average delay constraints,” *Proc. Allerton Conference on Commun.,Control, Comp., Monticello, IL.*, September 1999.

[14] W. Wu, A. Arapostathis, and S. Shakkottai, “Optimal power allocation for a wireless channel under heavy traffic approximation,” *The 38th Annual Conference on Information Sciences and Systems (CISS’04), Princeton University*, March 17 - 19 2004.

[15] A. Fu, E. Modiano, and J. Tsitsiklis, “Optimal energy allocation for delay-constrained data transmission over a time-varying channel,” *The 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), San-Francisco*, April 1 - 3 2003.

[16] ——, “Optimal energy allocation and admission control for communications satellites,” *The 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, June 23-27 2002.

[17] T. Holliday and A. Goldsmith, “Optimal power control and source-channel coding for delay constrained traffic over wireless channels,” *IEEE Int. Conf. on Comm. (ICC)*, 28 April - 2 May 2002.

[18] T. Holliday, “Cross-layer design of wireless networks,” *ISS Seminar, Stanford University*, March 12 2004.

[19] T. Holliday, A. Goldsmith, P. Glynn, and N. Bambos, “Distributed power and admission control for time varying wireless networks,” *IEEE Int. Symp. Inform. Theory (ISIT’04), Chicago, U.S.A.*, June 27 - July 2 2004.

[20] S. Adireddy and L. Tong, “Exploiting decentralized channel state information for random access,” *submitted to IEEE Trans. on Inform. Theory*, November 2002.

[21] G. Caire and D. Tuninetti, “The throughput of hybrid-ARQ protocols for the Gaussian collision channel,” *IEEE Trans. on Info. Theory*, vol. 47, no. 5, pp. 1971–1988, July 2001.

[22] M. Medard, S. P. Meyn, J. Huang, and A. J. Goldsmith, “Capacity of time-slotted ALOHA systems,” *IEEE Int. Symp. Inform. Theory (ISIT’00), Sorrento, Italy*, p. 407, June 25-30 2000.

[23] M. Sharif and B. Hassibi, “Delay guarantee versus throughput in broadcast fading channels,” *IEEE Int. Symp. Inform. Theory (ISIT’04), Chicago, U.S.A.*, June 27 - July 2 2004.

[24] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, “Throughput-delay trade-off in energy constrained wireless networks,” *IEEE Int. Symp. Inform. Theory (ISIT’04), Chicago, U.S.A.*, June 27 - July 2 2004.

[25] M. Medard, J. Huang, A. J. Goldsmith, S. P. Meyn, and T. P. Coleman, “Capacity of time-slotted ALOHA packetized multiple-access systems over the AWGN channel,” *IEEE Trans. on Wireless communications*, vol. 3, no. 2, pp. 486–499, March 2004.

[26] S. Sesia, G. Caire, and G. Vivier, “On the scalability of H-ARQ systems in wireless multicast,” *IEEE Int. Symp. Inform. Theory (ISIT’04), Chicago, U.S.A.*, June 27 - July 2 2004.

[27] ——, “Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes,” *to appear in IEEE Trans. on Commun.*, 2004.

[28] E. Yeh and A. Cohen, “Throughput optimal power and rate control for queued multiaccess and broadcast communications,” *IEEE Int. Symp. Inform. Theory (ISIT’04), Chicago, U.S.A.*, June 27 - July 2 2004.

[29] S. Shamai (Shitz), “A broadcast approach for the multiple-access slow fading channel,” *IEEE Int. Symp. Inform. Theory (ISIT’00), Sorrento, Italy*, p. 128, June 25-30 2000.

[30] ——, “A broadcast strategy for the Gaussian slowly fading channel,” *IEEE ISIT’97, Ulm Germany*, p. 150, June 29–July 4 1997.

[31] S. Shamai (Shitz) and A. Steiner, “A broadcast approach for a single user slowly fading MIMO channel,” *IEEE Trans. on Info, Theory*, vol. 49, no. 10, pp. 2617–2635, Oct. 2003.

[32] A. Steiner and S. Shamai (Shitz), “Multi-layer broadcasting for a faded MIMO channel,” *submitted to IEEE Trans on Info. Theory*, Oct. 2003.

[33] ——, “Multi-layer broadcasting in a MIMO channel,” *The 38th Annual Conference on Information Sciences and Systems (CISS’04), Princeton University*, March 17 - 19 2004.

[34] L. Ozarow, S. Shamai (Shitz), and A. Wyner, “Information theoretic considerations for cellular mobile radio,” *IEEE Trans. Veh. Tech.*, vol. 43, no. 2, pp. 359–378, May 1994.

[35] E. Biglieri, J. Proakis, and S. Shamai (Shitz), “Fading channels: Information theoretic and communication aspects,” *IEEE Trans. on Info. Theory*, vol. 44, no. 6, pp. 2619–2692, October 1998.

[36] T. Cover, “Broadcast channels,” *IEEE Trans. on Info. Theory*, vol. 18, no. 1, pp. 2–14, Jan. 1972.

[37] T. Cover and J. Thomas, *Elements of Information Theory*. New-York: Wiley, 1991.

[38] D. Tse, “Optimal power allocation over parallel Gaussian broadcast channels.” *Proceedings of International Symposium on Information, Ulm Germany, http://degas.eecs.berkeley.edu/d̃tse/pub.html*, p. 27, June 1997.

[39] L. Li and A. Goldsmith, “Capacity and optimal resource allocation for fading broadcast channels: Part I: Ergodic capacity and part II: Outage capacity,” *IEEE Trans. on Inform. Theory*, vol. 47, no. 3, pp. 1083–1127, March 2001.

[40] M. Sajadieh, F. R. Kschischang, and A. Leon-Garcia, “Analysis of two-layered adaptive transmission systems,” *IEEE, 46th Vehicular Technology Conference (VTC’96), Atlanta, Georgia*, pp. 1771–1775, April 28-May 1 1996.

[41] P. Schramn, “Multilevel coding with independent decoding on levels for efficient communications on static and interleaved fading channels,” *Proceedings of IEEE PIMRC’97*, pp. 1196–1200, Helsinki, Findland 1997.

[42] D. Schill and J. Huber, “On hierarchical signal constellations for the Gaussian broadcast channel,” *Proc. International Conference on Telecommunications (ICT’98), Porto Carras, Greece*, pp. 34–38, 21-25, June 1998.

[43] D. Schill, D. Yuan, and J. Huber, “Efficient broadcasting using multilevel codes,” *IEEE Inform. Theory Workshop (ITW’99), Metsovo, Greece*, p. 72, June 27 - July 1 1999.

[44] E. Yeh and A. Cohen, “Information theory, queueing, and resource allocation in multi-user fading communications,” *The 38th Annual Conference on Information Sciences and Systems (CISS’04), Princeton University*, March 17 - 19 2004.

[45] J. G. Dai and C. Li, “Stabilizing batch-processing networks,” *Operational Research INFORMS*, vol. 51, no. 1, pp. 123–136, JanuaryFebruary 2003.

[46] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, New-Jersey: Perentice-Hall, 1989.

[47] L. Kleirock, *Queueing Systems Volume 2: Theory*. New-York: John Wiley, 1975.

[48] D. Daley and C. Trengrove, *Bounds for mean waiting times in single server queues: A Survey*. Australian National University: Statistics Department (IAS), 1977.

[49] Y. Liu, K. Lau, C. Takeshita, and M. Fitz, "Optimal rate allocation for superposition coding in quasi-static fading channels," *IEEE ISIT'02, Lausanne, Switzerland*, p. 111, June 30 - July 5 2002.

[50] L. Kleirock, *Queueing Systems Volume 1: Theory*. New-York: John Wiley, 1975.

[51] A. J. Viterbi, "Very low rate conventional codes for maximum theoretical performance of spread-spectrum multiple-access channels," *IEEE Journal on Selected Areas in Communications*, vol. 8, no. 4, pp. 641–649, May 1990.

[52] I. Geldfand and S. Fomin, *Calculus of Variations*. Mineola, New-York: Dover Publications, Inc., 1991.

[53] E. Yeh and A. Cohen, "An inter-layer view of multiaccess communications," *IEEE Int. Symp. Inform. Theory (ISIT'02), Laussane, Switzerland*, p. 112, June 30 - July 5 2002.

[54] X. Qin and R. Berry, "Exploiting multiuser diversity in wireless ALOHA networks," *Proc. Allerton Conf. on Communication, Control and Computing, (Allerton, IL)*, October 2001.

[55] M. Stege and G. Fettweis, "Successive interference cancellation and ixs implications for the design of layered MIMO-algorithms," *IEEE ISIT'03, Yokohoma, Japan*, p. 811, June 29 - July 4 2003.

[56] S. Simon and A. Moustakas, "Optimizing MIMO antenna systems with channel covariance feedback," *IEEE Jnl. on Sel. Areas in Comm.*, vol. 21, pp. 406–417, Apr. 2003.

[57] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. on Telecomm.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.

[58] S. Verd$\acute{u}$ and S. Shamai (Shitz), "Spectral efficiency of CDMA with random spreading," *IEEE Trans. on Info. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.