

On the transmission of bursty sources

Stéphane Musy Emre Telatar
School of Computer and Communication Sciences, EPFL
CH-1015 Lausanne, Switzerland
Email: {stephane.musy,emre.telatar}@epfl.ch

Abstract—Traditionally, the bursty nature of data sources is not taken in consideration by information theory. Random arrival times typically are assumed to be smoothed out by appropriate source coding, rendering any meaningful analysis of the end-to-end delay impossible. On the other hand, network theory directly treats these issues, but over-simplifies the channel model. Particularly, the issues of noise and interference are ignored and no sophisticated coding is allowed. In this paper, we introduce a framework in which some aspects of both sides are incorporated. This results in the formulation of new scheduling problems. In simple settings, we are able to characterize and analyze delay optimal policies.

I. INTRODUCTION

The bursty nature of data sources (due to random message arrivals) is mainly addressed by network theory. Within, the random arrival of packets is directly treated via resource allocation and scheduling algorithms. Powerful tools have been developed to analyze network layer quantities, such as end-to-end delay. But, the channel model is over-simplified. Particularly, the issues of noise and interference are ignored and no inter-packet coding is allowed. On the other hand, traditional information theory provides accurate models for the transmission process, but typically no consideration is made for the random arrival of the messages. Random arrival times are assumed to be smoothed out by appropriate source coding. This renders any meaningful analysis of the end-to-end delay impossible. To have a genuine understanding of the transmission of bursty sources, we need to combine these two approaches, as highlighted by Ephremides and Hajek [1].

A particular case of interest is the multiaccess communication system. In this setting, source burstiness becomes a fundamental issue [2]. In the late 90's, work has been done taking a lead from [3] to combine these two perspectives, essentially by borrowing tools from queueing theory and information theory. Recently, the problem of resource allocation (power control and rate allocation) in multiaccess communication, has shown to be essential in the characterization of the optimal values for quality-of-service measures like packet throughput and delay [4], [5], [6].

All these works seek to minimize the average delay of packets with an appropriate rate allocation policy, where the control space is given by the multiaccess capacity region. Furthermore, it is assumed that each packet can be sent at the maximum achievable rate, and each packet is considered independently of the others at a given transmitter. In this paper

The work presented in this paper was partially supported by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322.

we shall focus on the single user case, and introduce a simple minded framework which allows to take into consideration, via channel latency, certain simple models of correlation between packets and the possibility that the packets are too short to perform capacity achieving coding.

In the next section, we introduce our framework and formulate the problem. In Section III, we treat a simple channel model in which we can show that the optimal scheduling policy is a threshold based strategy. Then, in Section IV, we analyze the case where the channel latency is affine. We give lower and upper bounds to the average packet delay a policy can achieve. Finally, in Section V, we analyze and characterize the optimal policies in the affine latency case, when there are no more packet arrivals in the system. An interesting symmetry property of those policies is exhibited.

II. FRAMEWORK AND PROBLEM FORMULATION

We keep the conventional source-channel-destination model, the system is composed of a transmitter driven by a bursty source, a channel and a receiver. The transmitter is separated in two components, a queue storing the packets and a scheduler deciding when the next transmission should start and which packets have to be sent (see Fig. 1). We assume that the packets have the same length and that the arrival times in the transmitter queue follow a stochastic process of rate λ .

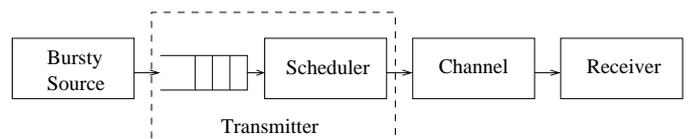


Fig. 1. System Model

For the purpose of analyzing the end-to-end delay appearing in the communication system, we characterize the channel by its latency $\mathcal{T}(n)$. This is the overall transmission delay a *superpacket* containing n packets will suffer when it is encoded and sent through the channel.¹ The latency function can be derived using tools from information theory. We make these two general assumptions on the latency $\mathcal{T}(n)$:

- 1) $\mathcal{T}(n+1) \geq \mathcal{T}(n)$ (increasing)
- 2) $\mathcal{T}(n+m) \leq \mathcal{T}(n) + \mathcal{T}(m)$ (sub-additive)

The first assumption is evident: sending more packets takes more time. The second one, comes from the use of coding.

¹This definition supposes that all the packets are decoded with arbitrary small probability of error. We can envisage other definitions in which a tradeoff is made between delay and probability of error.

We can relate the latency to the source entropy rate h and the channel capacity c (per packet), by: $\mathcal{T}(n) \geq nh/c$. Furthermore, the second assumption ensures that $\lim_{n \rightarrow \infty} \frac{\mathcal{T}(n)}{n} = \inf_n \frac{\mathcal{T}(n)}{n}$. And, the source-channel coding theorem implies that $\lim_{n \rightarrow \infty} \frac{\mathcal{T}(n)}{n} = \frac{h}{c}$. Thus, the quotient of the source entropy rate over the channel capacity describes the asymptotic behavior of the latency.

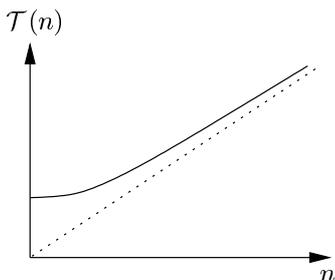


Fig. 2. Example of channel latency curve

Once \mathcal{T} and λ are given, the problem is to derive policies that map each state of the transmitter queue to a number of packets to send. We will look at scheduling policies that are optimal in the sense of achieving the lowest average packet delay (delay-optimal). Here delay is to be understood as the end-to-end packet transmission duration, which is the addition of the waiting time (in the transmitter queue) and the transmission time.

In the assumption that each packet can be sent at the maximum achievable rate of the channel, the latency is a linear function, and it is easy to show that a delay-optimal policy should send each packet successively. Nonetheless, for general latency curves, it seems that this problem is difficult. In the sequel, we study two simple but representative latencies, for which we are able to indicate optimal or good policies.

III. CONSTANT LATENCY

In this section, we consider the extreme situation where $\mathcal{T}(n) = D$, this represents a channel over which it takes a fixed amount of time D to transmit packets, independent of the number of packets sent. Though channels with a constant latency have a unlimited capacity, and therefore are not realistic, they are interesting as a limiting case of channels having a large bandwidth and a non-negligible setup time.

A. Optimal Policy

Under the assumption of Poisson arrivals, we will prove that the delay-optimal scheduling policy for channels with a constant latency will be a threshold strategy. That is, whenever the transmitter is ready he should send all packets waiting in his queue, only if there are more than a given number, otherwise, he should wait for more packets to arrive. This is the subject of the following proposition and theorem. Before enunciating them, we give a definition: we denote by a *transmission phase* a period of time in which the transmitter sends one or several packets, in our case it would always be of duration D .

We state here a straightforward proposition which is valid for any kind of arrival time statistics.

Proposition 3.1: In a single user communication system with a constant channel latency, a delay-optimal policy should send all the packets waiting in the transmitter queue, whenever a transmission phase is started.

Hereafter, we use Little's law to relate the average packet delay \bar{D} , with the time-average number of packets in the system \bar{N} :

$$\lambda \bar{D} = \bar{N}.$$

In order to prove our theorem, we need two simple lemmas.

Lemma 3.2: For a channel with a constant latency, minimizing the average packet delay is equivalent to minimizing the average waiting time.

Lemma 3.3: In a single user communication system with Poisson packet arrivals, the state of the transmitter queue is only given by the number of packets contained in it.²

Now, we are able to prove the optimality of threshold based strategies.

Theorem 3.4: In a single user communication system with a constant channel latency and Poisson packet arrivals, the delay-optimal scheduling policy is a threshold strategy.

Proof: The proof rely on a sample path analysis with the use of an interchange argument. This is a standard proof technique in scheduling theory (see, e.g., [7]). We will show that, if it is optimal for the transmitter to send when there are n packets waiting in the queue, the transmitter should also decide to send when there are $n + m$ packets waiting in the queue $\forall m \in \mathbb{N}$. (Since, to avoid infinite delay, any policy should decide to send when there are a certain number of packets in the queue, the preceding will prove that a threshold based strategy is optimal.)

By Lemma 3.3, we let the transmitter make a decision (send or wait) only by looking at the number of packets waiting in his queue. Now, let us assume that in the optimal policy, the transmitter decides to send whenever n packets are in his queue. This means, that averaging over all possible future paths the cost (in term of delay) of sending when there are n packets is less than the cost of waiting. In the following, we will consider two different scenarios, one where the system starts with n packets in the transmitter queue and another where the system starts with $n + m$ packets. Furthermore, we assume that we have the same realization of the future arrival process in both scenarios. We can use this coupling argument, since the arrivals are Poisson, and thus the future arrival times are independent of the current number of packets in the queue.

In both scenarios, we want to compare the average packet delay, when the transmitter decides to send immediately and when he decides to wait for a certain amount of time. By the Little's law and Lemma 3.2 it is sufficient to compare the integral over time of the number of packets waiting in the queue. For this, we let the transmitters, in both cases, use the same optimal policy (whatever it is) after the first packet

²Notice that with other arrival processes, the state may have to include arrival times of the packets as well as the current time.

arrival. Denote by N_s the integral over time of the number of packets waiting in the queue when the transmitter decides to send (shaded area in Fig. 3), and by N_w the same quantity when the transmitter decides to wait (shaded area in Fig. 4), for the first scenario. Similarly, denote by N'_s and N'_w these quantities for the second scenario.

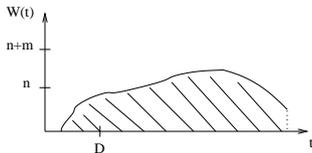


Fig. 3. Execution of the system when the transmitter sends immediately. In the figure, $W(t)$ represents the number of packets waiting in the queue at time t .

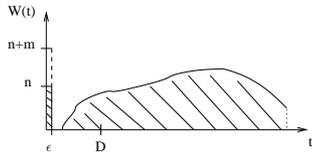


Fig. 4. Execution of the system when the transmitter waits a certain amount of time ϵ . In the figure, $W(t)$ represents the number of packets waiting in the queue at time t .

Note that, in the case when the transmitter decides to send immediately, Proposition 3.1 and the fact that the arrival times in both realizations are identical, imply that the transmitter queues in both scenarios will reach the same state after the transmission of the first packets. Thus, we have $N'_s = N_s$ (equal to the shaded area in Fig. 3). Furthermore, in the case where the transmitter chooses to wait, there are more packets waiting in the queue in the second scenario than in the first, so the integral number of packets could only increase implying $N'_w \geq N_w$.

Thereby, we have seen that $N_s = N'_s$ and $N_w \leq N'_w$ under the assumption of identical future arrivals. Since we know that $E[N_s] \leq E[N_w]$ (expectations are taken over all future paths) and since the future arrivals in each scenario have the same distribution, $E[N_s] = E[N'_s]$ and $E[N_w] \leq E[N'_w]$ which implies $E[N'_s] = E[N_s] \leq E[N_w] \leq E[N'_w]$. Thus, we see that it is also better to transmit when we have $n + m$ packets. ■

It is not too difficult to obtain an analytical expression of the average number of packets in the system related to the threshold, which allows us to compute the optimal threshold. This optimum threshold can be shown to be between $\frac{\lambda D}{2}$ and λD , where λD corresponds to the average number of packet arrivals during a transmission phase.

IV. AFFINE LATENCY

Here, we will consider a more realistic channel model by letting the latency to be affine, i.e., $\mathcal{T}(n) = D + nk$ for some positive numbers D and k . This model gives a tradeoff between the previous extreme case, and the linear behavior. Moreover, it is a good formulation for channel having a non-negligible setup-time. Observe that such a channel has a capacity of $\frac{1}{k}$ packets per unit of time. Hence, in the following we will only consider arrival rates such that $\lambda k < 1$.

We state here, without a proof, a general proposition about optimal policies:

Proposition 4.1: For a channel with an affine latency, packets that are present before the beginning of a transmission

phase and which are not sent, could only increase the end-to-end delay of the future packets.

This proposition does not mean that an optimal strategy should send all packets waiting in the queue, since in the affine case the packets sent increase also the delay of the other packets in the transmission phase. However, under the assumption of Poisson arrivals, it is tempting to say that such a strategy will perform well. The following paragraphs will establish lower and upper bounds on the performance of optimal policies.

A. Lower Bound on Average Delay

The average end-to-end delay \bar{D} can be decomposed as the sum of the average waiting time \bar{D}_w in the transmitter queue and the average transmission time \bar{D}_t . In this section, we start by deriving a lower bound on the average transmission time for any arrival process of rate λ , then we give a lower bound on the average end-to-end delay, by assuming Poisson arrivals. In this purpose, we use an argument similar to that in [4]. Let us look at our system as a single-server queue with a variable service rate depending on the number of packets being transmitted. Indeed, for any number of packets transmitted n , the channel latency defines a service rate $R(n) = \frac{n}{\mathcal{T}(n)}$.

Now, suppose that packets of same length arrive in the system at a rate of λ , and let p_n denote the long term fraction of time during which n packets are transmitted under a service policy. Thus, the time average number of packets being transmitted (served) is $\bar{N}_t = \sum_n n p_n$, and by Little's law the average transmission time satisfies $\lambda \bar{D}_t = \bar{N}_t$.

With the preceding observation, the long term average service rate can be expressed as $\sum_n R(n) p_n$. For stability we need $\lambda \leq \sum_n R(n) p_n$, so we have:

$$\bar{N}_t \geq \inf \left\{ \sum_n n p_n : \sum_n R(n) p_n \geq \lambda \right\},$$

or equivalently,

$$\begin{aligned} \lambda &\leq \sup \left\{ \sum_n R(n) p_n : \sum_n n p_n \leq \bar{N}_t \right\} \\ &= \sup \{ E[R(N)] : E[N] \leq \bar{N}_t \}, \end{aligned}$$

where the supremum is over all non-negative integer valued random variable N . Let $t : [0, \infty) \rightarrow [0, \infty)$ be obtained by linearly interpolating of the channel latency function \mathcal{T} ; $t(x) = D + xk$. Then, we can define $r(x) = \frac{x}{t(x)}$, which is a concave, increasing function. Thus, by relaxing the integer restriction on N , and using Jensen's inequality

$$\lambda \leq r(\bar{N}_t).$$

Combined with the Little's law this yields: $\bar{D}_t \geq r^{-1}(\lambda)/\lambda$, where r^{-1} is the inverse function of r . Letting \hat{n} be such that $\hat{n} = \lambda t(\hat{n})$, we can write $\bar{D}_t \geq \frac{\hat{n}}{\lambda} = t(\hat{n})$.

With $\mathcal{T}(n) = D + nk$, we get $\hat{n} = \frac{\lambda D}{1 - \lambda k}$, and thus

$$\bar{D}_t \geq \frac{D}{1 - \lambda k}.$$

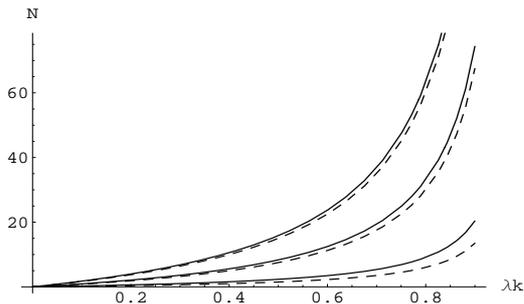


Fig. 5. Upper and lower bounds to \bar{N} for different values of the ratio D/k , related to λk . The curves are for D/k values of 1, 5 and 10 (from right to left).

This lower bound holds for every channel having the property that $r(x)$ is a concave function.

We will now derive a lower bound for the end-to-end delay \bar{D} , by restricting our focus on Poisson arrivals of rate λ . Let us decompose the time average number of packets in the system \bar{N} , as $\bar{N} = \bar{N}_w + \bar{N}_t$, the sum of the time average number of packets waiting in the queue and the time average number of packets being transmitted. As before, we let p_n denote the long term fraction of time during which n packets are transmitted under a service policy. Thus, we can write $\bar{N}_t = \sum_n n p_n$. Assuming Poisson arrivals, \bar{N}_w is lower bounded by $\bar{N}_w \geq \sum_n \frac{\lambda T(n)}{2} p_n$, which is obtained by only counting the average number of packets present in the queue, at the end of a transmission phase. With the constraint of stability $\lambda \leq \sum_n R(n) p_n$, we get:

$$\bar{N} \geq \inf \left\{ \sum_n n p_n + \sum_n \frac{\lambda T(n)}{2} p_n : \sum_n R(n) p_n \geq \lambda \right\}.$$

Since, $T(n)$ is an affine function of n , the set of $\{p_n\}$ which minimizes $\sum_n n p_n$, will also minimize $\sum_n \frac{\lambda T(n)}{2} p_n$. Using the previous result we can directly write

$$\bar{N} \geq \hat{n} + \frac{\lambda t(\hat{n})}{2} = \frac{3}{2} \hat{n}.$$

Which gives the following lower bound on the end-to-end delay:

$$\bar{D} \geq \frac{3}{2} t(\hat{n}) = \frac{3}{2} \frac{D}{(1 - \lambda k)}.$$

B. Threshold Policy

Using a policy with a threshold equal to one, i.e., which sends all the packets as soon as possible, we obtain an upper bound on the average number of packets in the system, for the optimal policy.

This upper bound as well as the lower bound obtained previously are shown in Fig. 5. One can observe that, in certain regimes, the upper and lower bounds are tight. Particularly, in low rate regime and for large values of D/k , we see that not much could be gain using more elaborate strategies.

Note that the strategy sending packets one after the other, performs poorly when the ratio D/k becomes larger than 1: indeed, for such a strategy the average packet delay is

proportional to $\frac{(D+k)}{1-\lambda(D+k)}$, and we see that this term diverges when $\lambda(D+k)$ approaches 1.

V. NO MORE ARRIVALS

Motivated by the fact that threshold policies perform well under affine latency, we study the scheduling problem consisting of transmitting a set of packets, when no more packets arrive. That is, we want to find the sequence of packets transmission that minimizes the average time a packet spends in the system.

A. Model

Let N denote the number of packets to transmit. A transmission policy P is a sequence of positive integers $\{n_i, i = 1, \dots, m\}$, with $\sum_i n_i = N$, where n_i is the number of packets sent at the i th stage over a total of m transmission stages. The transmission time is then a function of the number of packets sent: $T(n_i) = n_i + d$, where d is a constant denoting the intrinsic delay of a transmission.³ The average of the times the packets spend in the system under a given policy P can be found as

$$\bar{D}(P) = \frac{1}{N} \sum_{i=1}^m T(n_i) (N - \sum_{j<i} n_j).$$

For a given N and d , we look for a policy P which minimizes \bar{D} , such a policy will be called optimal.

B. Optimal Policies

A naive way to find the optimal strategy is to compare all possible policies, for a given d , and determine which one minimizes the cost function. Of course, such an approach is not efficient for large values of N , as the following proposition shows the number of possible policies grows exponentially with N [8].

Proposition 5.1: Let \mathcal{P} represent the ensemble of all possible policies, then

$$|\mathcal{P}| = 2^{N-1}.$$

However, an easy way to list all optimal policies that send up to N packets exists. Observe that, among all policies which begin by sending k packets, finding the best one is equivalent to find the optimal policy transmitting $N - k$ packets. Now, suppose that we know all optimal policies for $n < N$. To determine the optimal policy that sends N packets, we just have to compare N different strategies (one for all $k < N$). Repeating this argument, we see that with only $O(N^2)$ comparisons we can list all optimal policies that have up to N packets to transmit [8].

The figures show at the top of the next page are examples of optimal policies. In these pictures, the $\{n_i\}$ are represented by the number of squares in each row.

From these examples, we detect an interesting symmetry between optimal policies for d and $\frac{1}{d}$: in a sense explained below these strategies are conjugates. Actually, this symmetry holds for every optimal policy as shown in the next section.

³Note that this choice of $T(n_i)$ takes in consideration all affine functions of the form $n_i k + D$.

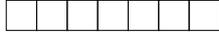


Fig. 6. Example of optimal policy for $d \geq N$.

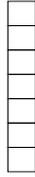


Fig. 7. Example of optimal policy for $\frac{1}{d} \geq N$.

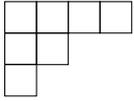


Fig. 8. Example of optimal policy for $d \in (\frac{1}{N}, N)$.

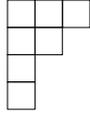


Fig. 9. Example of optimal policy for $\frac{1}{d} \in (\frac{1}{N}, N)$.

C. Properties of Optimal Policies

First, let us state a general criterion on the local optimality of a policy.

Proposition 5.2: Let $P = \{n_1, n_2, \dots, n_m\}$ be an optimal policy, then $n_{i+1} \leq n_i$ for all $1 \leq i < m$. This means that, in an optimal policy, the sequence of packets sent at each transmission is non-increasing.

Proof: Let $P = \{n_1, n_2, \dots, n_m\}$ be a policy such that $n_{i+1} > n_i$ for a given i . Denote by $R_i = N - \sum_{j < i} n_j$, the number of remaining packets at stage i . With $R_{i+1} = R_i - n_i$, we can write the cost due to the successive stages i and $i+1$ as

$$R_i(n_i + d) = R_i n_i + R_i d$$

$$(R_i - n_i)(n_{i+1} + d) = R_i n_{i+1} - n_i(n_{i+1} + d) + R_i d.$$

Since $R_{i+2} = R_i - (n_i + n_{i+1})$, n_i and n_{i+1} will interact with the next stages only through their sum. Thus, the interchange of n_i and n_{i+1} only affect the stage i and $i+1$. Now, observe that a policy $P' = \{n'_1, n'_2, \dots, n'_m\}$ equal to the policy P with the exchange of n_i and n_{i+1} , such that $n'_{i+1} = n_i \leq n'_i = n_{i+1}$, have its total cost decreased by $(n_{i+1} - n_i)d$. ■

In order to formulate the symmetry relation seen previously, we look at conjugate policies and their properties.

Definition 5.1: Given a policy $P = \{n_i, i = 1, \dots, m\}$, we define its *conjugate* policy $P^* = \{n_k^*, k = 1, \dots, l\}$ by letting $n_k^* = |\{j : n_j \geq k\}|$ and $l = \max n_i$.

In the example shown by Fig. 10, the $\{n_i\}$ are represented by the number of squares in each row, and the $\{n_k^*\}$ by the number of squares in each column. Therefore, a policy and its conjugate can be seen as two different perspectives (horizontal and vertical) of the same object.

From this picture, we see that summing the squares from left to right or from top to down will give the same result, i.e., $\sum n_i = \sum n_k^* = N$. Using similar relations, we are able to

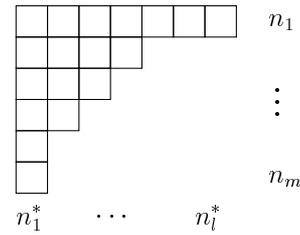


Fig. 10. Representation of a policy.

prove the following theorem on the symmetry relation between optimal policies. A proof of this statement can be found in [9].

Theorem 5.3: If a policy P is optimal for a delay d , then its conjugate policy P^* is optimal for a delay of $\frac{1}{d}$.

VI. CONCLUSION

We introduced a new measure of channel features that allows to incorporate some notions related to information theory in a scheduling problem. In this setting, we look for policies that minimize the average packet delay. For constant channel latencies, and when the packet arrivals follow a Poisson process, we showed that threshold based strategies are optimal. When the channel latency is an affine function, we established lower and upper bounds on the performance a policy can achieve. These bounds reveal that threshold policies are good in certain regimes. Finally, we analyzed the no more arrivals case with an affine latency. This turns out to be an interesting combinatorial problem, for which we exhibited a symmetry relation between optimal policies.

The formulation described here is perhaps simplistic in its approach to modeling the physical layer, and in this paper we have only discussed simple cases of latency function. Nevertheless, we believe that this framework can give new intuition to questions that fall in the intersection of information theory and network theory.

REFERENCES

- [1] A. Ephremides and B. Hajek, "Information Theory and Communication Networks: An Unconsummated Union," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2416-2434, October 1996.
- [2] R. G. Gallager, "A Perspective on Multiaccess Channels," *IEEE Trans. Inf. Theory*, vol. 31, no. 2, pp. 124-142, March 1985.
- [3] E. Telatar and R. G. Gallager, "Combining Queueing Theory with Information Theory for Multiaccess," *IEEE J. Sel. Areas Comm.*, vol. 13, no. 6, pp. 963-969, August 1995.
- [4] S. Raj, E. Telatar and D. Tse, "Job Scheduling and Multiple Access," *DIMACS Series in Disc. Math. and Th. Comput. Sci.*, vol. 66, 2004.
- [5] E. Yeh, "Delay-Optimal Rate Allocation in Multiaccess Communications: A Cross-Layer View," *Proc. 2002 Int. Workshop On Multimedia Signal Processing*, St Thomas, US Virgin Islands, December 2002.
- [6] E. Yeh and A. Cohen, "Information Theory, Queueing, and Resource Allocation in Multi-user Fading Communications," *Proc. 2004 Conf. Inf. Sci. and Syst.*, Princeton, NJ, March 2004.
- [7] D. Towsley, "Application of Majorization to Control Problems in Queueing Systems," chap. 14 in *Scheduling Theory and its Applications*, Wiley, 1995.
- [8] V. Pasquier, "Information Theoretic Packet Scheduling for Single-Server Queues," Master thesis, Information Theory Laboratory, EPFL, September 2003.
- [9] S. Musy, "Packet Scheduling for Communication Systems with Cost of Transmission," Information Theory Laboratory, EPFL. Available: <http://lthiwww.epfl.ch/~smusy/publications.html>