A Survey of Recent Results on Real-Time Wireless Networking

I-Hong Hou
CSL and Department of Computer Science
University of Illinois
Urbana, IL 61801, USA
ihou2@illinois.edu

P. R. Kumar
CSL and Department of ECE
University of Illinois
Urbana, IL 61801, USA
prkumar@illinois.edu

***Abstract*—Over the past two years there has emerged a new approach to solve the problem of delivering required throughputs of packets that meet hard deadlines over heterogeneous unreliable channels. These recent results provide a contract for flows with throughput as well as delay constraints. These contracts have two desirable properties: the contracts can be supported by the wireless network, and, further, the contracts are appropriate enough that applications can define their requirements through them.**

The theory provides admission control algorithms for deciding when flows with throughput-cum-deadlines can be satisfied, as well as simple scheduling algorithms for doing so. The theory also extends well to various arrival patterns for packets, fading models, and rate adaptation schemes, as well as broadcast. It can also be generalized to optimize the service of elastic flows that have utilities based on the throughput provided. Further, it can be used in an incentive compatible way for strategic auctions. The results of the above theoretical framework are surprising, elegant and simple for inelastic as well as elastic flows, in terms of admission control for the form and scheduling for both.

We provide an account of this emerging theory.

I. Introduction

There are increasing demands for using wireless networks to serve applications that require delay guarantees. Such applications include VoIP, video streaming, real-time surveillance, networked control, etc. One common characteristic of these applications is that they have a strict deadline associated with each packet. Each packet needs to be delivered before its deadline, or it expires and is no longer useful for its application. On the other hand, while such applications may tolerate a small portion of their packets missing their deadlines, they still require a specific *timely throughputs*, which is defined as the throughputs of packets that are delivered before their deadlines, in order to maintain their performance.

Serving such applications is especially challenging in wireless networks. Wireless transmissions are subject to shadowing, fading, and interference from other transmissions. Thus, wireless transmissions are usually unreliable. Further, the channel reliabilities of different clients can be different, and can even vary over time.

There are of course several alternative ways in which to model a wireless network consisting of flows that have delay constraints. The difficulty has been that these formulations have by and large led to intractable analytical problems. The end result has been that essentially no progress has been made in this area. Recently, there has emerged a formulation that appears to provide a useful and tractable framework for the modeling, analyzing and designing real-time wireless communications. This formulation also appears to be generalizable in several directions to handle various additional features of problems while still providing tractable solutions and in some cases somewhat surprising answers. The purpose of this paper is to provide an account of these recent results. We will for the most part expound on the formulations and results, providing only brief insights into the results. For the latter, the reader is referred to the original references [3]–[8]. This framework is built on top of an analytical model that jointly considers the three important aforementioned challenges: a strict deadline for each packet, the timely throughput requirement specified by each client or application, and finally the unreliable and heterogeneous nature of wireless transmissions. An important feature is that this model is suitable for characterizing the needs of a wide range of applications, and allows each application to specify its individual demand. Thus the contracts that result from this framework are on the one hand supportable by protocols, and on the other usable by application designers.

The ensuing theory provides several important solutions for serving applications with both delay guarantees and timely throughput requirements. Its core is a sharp analytical characterization of when the set of demands of all the clients in a system are feasible under the limitations of their channel reliabilities. One can obtain a polynomial-time algorithm for admission control. Further, one can address the problem of packet scheduling. The theory results in an on-line scheduling policy that is feasibility optimal in the sense that it fulfills the demands of any set

This material is based upon work partially supported by USARO under Contract Nos. W911NF-08-1-0238 and W-911-NF- 0710287, NSF under Contracts CNS-1035378, CNS-0905397, CNS-1035340, and CCF-0939370, and AFOSR under Contract FA9550-09-0121.

of clients as long as the demands of the set of clients are feasible.

This framework can further be generalized in various directions of interest. It can be extended to address scenarios where the timely throughput requirements of clients are elastic. This can be formulated as a utility maximization problem. A bidding game, in which both clients and servers follow simple on-line strategies, then achieves the maximum total utility in the system.

The framework can also incorporate the usage of rate adaptation, where different clients use different transmission rates in order to guarantee error-free transmissions. Both the problems of packet scheduling and utility maximization can be addressed when rate adaptation is used. Further, one can also study the behaviors of selfish and strategic clients, and a truthful auction can be designed, under which strategic clients gain nothing by hiding their private preferences.

Finally, this framework can also be generalized to broadcast flows that require delay guarantees. In that case one can use network coding to enhance performance.

In this paper, we provide a survey of some of the main results. The rest of of the paper is organized as follows. Section II describes the basic framework and model. Section III addresses the problem of admission control. Section IV provides a feasibility optimal scheduling policy. Section V considers the problem of serving elastic traffic. Section VI shows that the framework can incorporate rate adaptation and address strategic behaviors of clients. Section VII extends the model to broadcast and discusses the usage of network coding. Finally, Section VIII concludes this paper.

II. The Basic Model

Consider a system with one access point (AP) and N wireless clients, labeled as $\{1, 2, \ldots, N\}$. Figure 1 illustrates an example of such a system. Each client generates a flow that has both a strict per-packet delay bound and timely throughput requirement. It is assumed that time is slotted, and the length of a time slot equals to the time needed to make a transmission between the AP and a client. The AP is in charge of scheduling transmissions in all time slots. When a flow involves uplink transmission, the AP uses some polling mechanism, such as the 802.11 PCF, to schedule such a transmission.

It is assumed that time slots are grouped into *intervals*, where an interval consists of T time slots. Each client generates one packet at the beginning of each interval, as shown in Figure 1. Each packet is associated with a deadline equal to T time slots. That is, packets generated at the beginning of an interval need to be delivered before the end of the interval. If a packet is not delivered by its deadline, it expires and is dropped from the system. Since only unexpired packets are candidates for being delivered, it follows that the delays of all successfully delivered packets are at most T time slots.

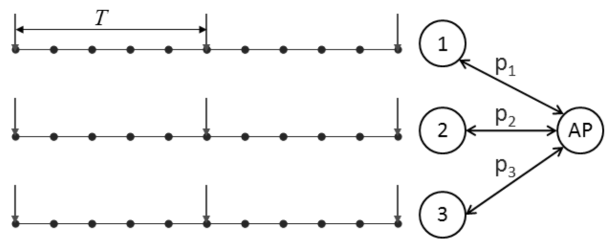

Fig. 1: An example that illustrates the system. The right half of the figure illustrates the topology of the system and the channel reliabilities between clients and the AP. The left half of the figure shows the timeline of each client, where each arrow indicates the arrival of a packet.

The unreliable and heterogeneous nature of wireless transmissions is a feature of the model. To begin with, we assume that whenever the AP schedules a transmission for client n, the transmission is successful with probability p_n. On the other hand, the packet is corrupted and the transmission fails with probability $1 - p_n$.

Since wireless transmissions are unreliable, it is not generally possible to deliver all packets before their deadlines. Therefore, we measure the performance of the service provided to a client by its *timely throughput*, which is defined as the long-term average number of packets that are delivered on-time for the client. More formally, the timely throughput of client n is $\liminf_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} 1($A packet of client n is delivered successfully in interval $k)$, where $1(\cdot)$ is the indicator function of the event.

Now we turn to the requirements of the clients. We suppose that each client n requires that its timely throughput to be at least q_n. We will say that a client n is *fulfilled* by a scheduling policy if, under the scheduling policy, its actual timely throughput is at least q_n, almost surely.

In summary, this model jointly considers three important features of real-time wireless networks: delay bounds on packets, timely throughput requirements of clients, and the unreliable and heterogeneous nature of wireless transmissions.

The above model was first proposed in [3]. In [5] and [6], it was shown how to extend it and incorporate scenarios where clients may generate traffic according to different traffic patterns, clients may have different delay bounds, channel qualities may subject to fading, and clients in the system use rate adaptation.

III. Admission Control

We now study the problem of admission control. This relates to the question: how can one determine verify whether it is theoretically feasible to fulfill all the requirements of all the clients in the system.

We begin with the following observations: Whether a client is fulfilled can be determined by evaluating the long-term average fraction of time that the AP spends in transmitting its packets.

Lemma 1: The timely throughput of client n is at least q_n with probability 1 if and only if the long-term average of the proportion of time slots that client n is transmitting is at least $w_n = \frac{q_n}{p_n T}$.

The above Lemma reduces the problem of feasibility to the following question: Can we ensure that the AP spends a proportion of time w_n transmitting the packets of client n for each n? Clearly, $\sum_{n=1}^{N} w_n$ cannot exceed one if the set of q_n's is to be feasible. Thus $\sum_{n=1}^{N} w_n \leq 1$ is a necessary condition for feasibility.

Is this sufficient? Unfortunately not. Consider a system where there are two clients, and the interval $T = 3$. Suppose that in the first slot the AP transmits the packet of the first client. This is successful with probability p_1. Now there are two remaining slots in the period, and one packet, that of client 2. Now the AP turns to client 2 and transmits it. This is successful with probability p_2. Now the AP is left with one slot where it has not packet at all to send. Thus it is forced to remain idle in this scenario. It is easy to see that the expected idle proportion is $p_1 p_2/3$. Therefore the sum $\sum_{n=1}^{N} w_n$ cannot ever be made equal to 1, for it cannot exceed $1 - \frac{p_1 p_2}{3}$. Hence we have obtained a stronger necessary condition.

One can generalize this. Let γ_n denote the geometrically distributed random variable that represents the number of attempts of client n before its packet is successfully delivered. Then $I(\{1, 2, \ldots, N\}) := \frac{E[\max\{0, T - \sum_{n=1}^{N} \gamma_n\}]}{T}$ is the unavoidable idle time when there are N clients to be served in T slots. Clearly, $\sum_{n=1}^{N} w_n \leq 1 - I(\{1, 2, \ldots, N\})$ is necessary condition for feasibility.

Is this condition sufficient? Unfortunately not, again. To see this consider the following example shown in [5].

Example 1: Consider a system with interval length $T = 3$, and two clients. The raliabilities for both clients are $p_1 = p_2 = 0.5$. Client 1 requires a timely throughput of $q_1 = 0.876$, while the timely throughput requirement of client 2 is $q_2 = 0.45$.

Now, we have:

$$\begin{aligned} w_1 &= 1.76/3, \\ w_2 &= 0.9/3, \\ I_{\{1\}} = I_{\{2\}} &= 1.25/3, \\ I_{\{1,2\}} &= 0.25/3. \end{aligned}$$

If we evaluate the condition for the subset of $S = \{1\}$, we find $w_1 = 1.76/3 > 1.75/3 = 1 - I_{\{1\}}$. This indicates that a system with only client 1 is not feasible, let alone a system with both clients 1 and 2. However, if we evaluate the condition for the set of all clients $\{1, 2\}$, we have $w_1 + w_2 = 2.66/3 < 2.75/3 = 1 - I_{\{1,2\}}$. Thus, merely evaluating the condition for the set of all clients is not enough. □

The difficulty in the above example that there is a subset of clients for which the busy time required to serve it, plus the unavoidable idle time were those to be the only clients available exceeds one. Thus we need to examine not only the whole set but also every one its subsets $S \subseteq \{1, 2, \ldots, N\}$. Let us define

$$I_S := \frac{E[\max\{0, T - \sum_{n \in S} \gamma_n\}]}{T},$$

where $\gamma_n \sim Geom(p_n)$. This is the unavoidable idle time when only the clients in set S need delivery. Clearly we need $\sum_{n=1}^{N} w_n \leq 1 - I(S)$ for *every* subset $S \subseteq \{1, 2, \ldots, N\}$, for feasibility. One may wonder why we need this condition. The answer is that the left hand side is monotone increasing in S, but so is the right hand side, since the idle time reduces as the client set is increased. Thus we cannot just only evaluate this when S is the set of all clients and conclude that it holds for all subsets S.

It turns out that the above condition is both necessary and sufficient, as stated in the following theorem.

Theorem 1: A set of clients is feasible if and only if $\sum_{n \in S} w_n \leq 1 - I_S$ holds for every subset S.

The problem of admission control is therefore solved by evaluating the above condition.

However, there is one further issue to consider – complexity. Evaluating this condition requires testing for all subsets, which requires checking 2^N any inequalities. However, it can be simplified resulting in total of N tests by the following theorem, resulting in a polynomial-time algorithm for admission control.

Theorem 2: Order the clients so that $q_1 \geq q_2 \geq \cdots \geq q_N$. Let S_k be the subset of clients $\{1, 2, \cdots, k\}$. The set of all clients is feasible if and only if $\sum_{n \in S_k} w_n \leq 1 - I_{S_k}$ for all k.

In summary, [3] provides a polynomial (in fact nearly linear) time algorithm for determining whether the requirements of a set of clients is feasible.

IV. Packet Scheduling

The policy is what is called a *largest debt first policy*. We need to define the concept of *debt*.

Definition 1: Let $c_n(t)$ be the number of successful transmissions of client n up to time t. The *debt* of client n at time t is defined as $(\frac{t}{T} \times q_n - c_n(t))/p_n$. It reflects how much the actual timely throughput is below its required timely throughput, q_n.

The largest debt first policy simply priority orders the clients at the beginning of each interval according to their debts and simply serves them in that order in the interval. Clients with larger debts get higher priority. In each time slot within the interval, the AP schedules the transmission for the highest priority client whose packet is not delivered yet.

Theorem 3: The largest debt policy satisfies the timely throughput requirements of all the clients whenever the set of clients is feasible.

The above result can be extended by defining a somewhat more general quantity called *pseudo-debt*. This provides a general approach to developing scheduling policies for other scenarios. Details are provided in [6] for problems where the channel reliabilities of clients change over time, and the case where rate adaptation is used and clients may have different delay bounds.

V. Serving Elastic Traffic

Up till now we have considered only the situation where the client' requirements are *inelastic*. That is, each client has a rigid requirement and specifies a timely throughput q_n that must be supported. We now treat the case where the client requirements are elastic. We suppose that each client has a *utility function*. Client n receives a utility $U_n(q_n)$ if its timely throughput is q_n. The total utility of the set of all clients is $\sum_{n=1}^{N} U_n(q_n)$. The goal of [7] is to find feasible $[q_n]$ to maximize the total utility in the system. This problem can be formally cast as the following optimization problem.

SYSTEM:

$$\text{Max} \sum_{i=1}^{N} U_i(q_i) \tag{1}$$

$$\text{s.t.} \sum_{i \in S} \frac{q_i}{T p_i} \leq 1 - I_S, \forall S \subseteq \{1, 2, \ldots, N\}, \tag{2}$$

$$\text{over } q_n \geq 0, \forall 1 \leq n \leq N. \tag{3}$$

Solving the above optimization problem directly is however difficult due to two reasons: First, the utility functions of different clients may be different, and the AP does not necessarily know the utility function of each client. Only the clients may know their own utilities, and they may not reveal their utility functions to the AP. Also there may be a large number of clients and the AP may not want to keep track of the utilities of all the clients. Second, there are exponentially many constraints for feasibility. Thus the optimization may be difficult to solve.

Instead of solving the problem directly, we show how the problem can be decomposed between the various clients and the AP. This is now a well known decomposition, see Kelly, Maulloo, and Tan [9] but we will show that in our case it has some surprising and, further, surprisingly simple consequences vis-a-vis knowledge of channel reliabilities, and the scheduling policy for the AP, respectively. Specifically, we decompose it into two subproblems, the **CLIENT**$_n$ subproblem and the **ACCESS-POINT** subproblem:

CLIENT$_n$**:**

$$\text{Max } U_n\left(\frac{\rho_n}{\psi_n}\right) - \rho_n \tag{4}$$

$$\text{over } 0 \leq \rho_n \leq \psi_n. \tag{5}$$

ACCESS-POINT:

$$\text{Max} \sum_{i=1}^{N} \rho_i \log q_i \tag{6}$$

$$\text{s.t.} \sum_{i \in S} \frac{q_i}{T p_i} \leq 1 - I_S, \forall S \subseteq \{1, 2, \ldots, N\}, \tag{7}$$

$$\text{over } q_n \geq 0, \forall 1 \leq n \leq N. \tag{8}$$

The two problems are decoupled by the notions of "price" announced by the AP to client n, and the amount "paid" by Client n to the AP. Let ψ_n denote the price per unit of timely throughput announced by the AP to client n. Given this, if the client pays ρ_n, then it obtains $\frac{\rho_n}{\psi_n}$ units of timely throughput. This brings it a utility $U_n(\frac{\rho_n}{\psi_n})$. However, it subtracts from this the amount it has paid, ρ_n, and obtains a "net benefit" $U_n(\frac{\rho_n}{\psi_n}) - \rho_n$, which Client n proceeds to maximize.

Now we turn to the AP. Once the clients have communicated their payments, it turns out that if the AP allocates the q_n's to maximize the objective function $\sum_{n=1}^{N} \rho_n \log q_n$, then it turns out that the resulting allocation of q_n's maximizes the overall sum of the utilities of all the clients. This result is established in [9].

How should the AP's optimization problem be solved? Note that this appears to be a complex problem because it has to be solved *subject to feasibility*, and that takes 2^N constraints. However, it turns out that it has a simple solution. All the AP has to do is give priority to clients for which the value of $\frac{\text{Number of time slots given to client n in [0,t]}}{\rho_n}$ is lowest, then that policy maximizes the AP's optimization problem. This is a rather simple policy!

However, there is yet another surprise. We note that neither the clients nor the AP need to know the channel reliabilities $\{p_n\}$! It turns out that the price automatically takes this into account.

It can be shown that if there is an equilibrium between the AP and the clients, which can be regarded as a "bidding game," then the overall price and amounts paid satisfy the property that the overall solution maximizes the total utility of the clients.

We note that the behavior of each client is consistent with its own interests. Also, it is interesting that there exists a simple on-line scheduling policy that solves **ACCESS-POINT**. Moreover, since this scheduling policy does not even need the information on p_n, it greatly reduces the overhead of channel probing and channel estimation.

VI. Incorporating Rate Adaptation and Addressing Strategic Behaviors

Many current wireless systems use rate adaptation in order to enhance transmission reliability. When rate adaptation is used, the AP effectively chooses for each client a modulation scheme that ensures a reliable communication with that client. Thus the AP uses different transmission rates for different clients, which results in different transmission times for packets for different clients.

We model this as follows. Let us suppose that a transmission between the AP and client n takes a total $s_n(t)$ time slots at time t. Note that this transmission time, $s_n(t)$, depends on t, which reflects the fact that channel qualities can change over time. We will assume that channel qualities change on a slower time scale than the length of an interval. Thus, $s_n(t)$ remains the same within an interval, but can change from interval to interval. It is assumed that when the AP chooses the appropriate rate for each client, all transmissions to the client are successful. Thus, there are no failed transmissions when rate adaptation is used.

We will also allow for different clients to have different delay bounds. Let us suppose that the delay bound of client n is τ_n time slots, where τ_n is smaller or equal to the length of the interval, T.

The system functions as follows. At the beginning of each interval, the AP decides a schedule for the interval by choosing an ordered subset of clients $S = \{m_1, m_2, \dots\}$ and transmits packets for them according to this ordering. The result is that the packet for a scheduled client m_l will be delivered at time $\sum_{n=1}^{l} s_n(t)$. Thus, we require that $\sum_{n=1}^{l} s_n(t) \le \tau_{m_l}$ for each scheduled client m_l in order to meet its deadline.

Similar to Section V, we suppose that each client has a utility function $U_n(\cdot)$, and receives utility $U_n(q_n)$ if its timely throughput is q_n. Let us consider the problem where the goal of the AP is to find a scheduling policy that maximizes the total utility in the system, $\sum_n U_n(q_n)$. In addition, we will also assume that each client n has a minimum timely throughput requirement $\underline{q}_n$, and the AP needs to ensure that $q_n \ge \underline{q}_n$. The latter assumption is made in order to enhance fairness and provide guarantees on minimum service for each client. The problem of maximizing the total utility can then be expressed as follows:

$$\text{Max } \sum_{n=1}^{N} U_n(q_n)$$
$$\text{s.t. Network dynamics and feasibility constraints,}$$
$$\text{and } q_n \ge \underline{q}_n, \forall n.$$

One particular challenge that arises is that "strategic" clients may lie about their utility functions in order to gain more service from the AP. Thus, in addition to solving the utility maximization problem, a mechanism that does not reward clients who lie about their true utility functions is also needed.

It turns out that we can solve this through an auction mechanism. In the auction, the AP announces a "discount" for each client at the beginning of each interval. Each client then makes a bid to the AP. Based on discounts and bids from clients, the AP selects some clients to schedule and charges them accordingly. The decisions on scheduling and charging clients are based on the Vickrey-Clarke-Groves (VCG) mechanism [1], [2], [10]. It can be shown that under this auction mechanism, all clients reveal their true utility functions. In addition, this mechanism also achieves the maximum total utility by adapting discounts for clients appropriately.

VII. Broadcasting Flows with Delay Bounds and Applying Network Coding

We can also extend this framework to encompass he problem of broadcasting flows with delay bounds. There are two distinct features vis-a-vis broadcasting. First, there can be multiple clients subscribing the same flow, and they may require different timely throughputs from that flow. Second, the AP does not obtain feedback information on whether a client receives a transmission successfully. This is because of the large overhead of gathering feedback from all clients. For this reason, ACKs are not implemented for broadcasting in most wireless mechanisms. This lack of feedback information introduces additional challenge to the problem of scheduling flows.

We can consider the problem of designing a scheduling policy where the scheduling decision for each time slot is based on the estimation of packet deliveries in previous time slots. In each time slot, the AP computes the *expected delivery debt*, which is defined to reflect the difference between the timely throughput requirement and the expected average number of actual packet deliveries, and the *marginal delivery probability*, which is the probability that a client can receive a packet that it has not received before from a flow if that packet is scheduled for transmission in the time slot, for each client and each flow. The *weighted marginal delivery probability* of a flow is then defined as the sum of products of its expected delivery debts and marginal delivery probabilities over all of its clients. AP then schedules the transmission for the flow who has the largest weighted marginal delivery probability. It turns out that this policy is feasibility optimal when network coding is not applied.

However, we can further consider the usage of network coding. Two coding schemes come to mind: XOR coding and linear coding. For each coding scheme, one can consider a scheduling policy. For XOR coding, a policy is designed by first pairing flows, and then computing the optimal number of transmissions for each of the raw packets of the two clients and the XOR-coded packet, under the restriction that the total number of transmissions of these three kinds of packets cannot exceed that of the two flows under the optimal policy when network coding is not applied. For linear coding, the proposed policy groups flows together, and then transmit linear combinations of packets within the same group. It turns out that both scheduling policies are feasibility optimal under mild restrictions, and they both offer better performances than the feasibility optimal policy without network coding. Thus, it follows that network coding can enhance performance even when strict per-packet deadline is enforced.

VIII. Conclusion Remarks

In this paper, we have provided a brief survey on an emerging theory for real-time wireless networks. This theory provides an analytical model that jointly considers three important factors for real-time wireless networks: strict per-packet deadlines, timely throughput requirements for clients, and the unreliable nature of wireless transmissions. We have presented solutions for admission control and packet scheduling. This theory can be extended to address the problem of serving clients with elastic traffic. It can also be extended to more complicated scenarios, such as the one where the channel qualities change over time, the AP applies rate adaptation, and clients are strategic. It is also useful for broadcasting flows with delay bounds and allows the usage of network coding.

We note that the results of Section III are from [3], the results of Section IV are from [3], [5], [6], the results of Section V are from [7], the results of Section VI are from [8], and the results of Section VII are from [4].

References

[1] E. Clarke. Multipart pricing of public goods. *Public Choice*, 11(1):17–33, Sep. 1971.

[2] T. Groves. Incentives in teams. *Econometrica*, 41(4):617–631, Jul. 1973.

[3] I-H. Hou, V. Borkar, and P.R. Kumar. A theory of QoS in wireless. In *Proc. of IEEE INFOCOM*, pages 486–494, 2009.

[4] I-H. Hou and P.R. Kumar. Broadcasting delay-constrained traffic over unreliable wireless links with network coding. to appear in Proc. of ACM MobiHoc 2011.

[5] I-H. Hou and P.R. Kumar. Admission control and scheduling for QoS guarantees for variable-bit-rate applications on wireless channels. In *Proc. of ACM MobiHoc*, pages 175–184, 2009.

[6] I-H. Hou and P.R. Kumar. Scheduling heterogeneous real-time traffic over fading wireless channels. In *Proc. of IEEE INFOCOM*, 2010.

[7] I-H. Hou and P.R. Kumar. Utility maximization for delay constrained QoS in wireless. In *Proc. of IEEE INFOCOM*, 2010.

[8] I-H. Hou and P.R. Kumar. Utility-optimal scheduling in time-varying wireless networks with delay constraints. In *Proc. of ACM MobiHoc*, 2010.

[9] F.P. Kelly, A.K. Maulloo, and D.K.H. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49:237–252, 1998.

[10] W. Vickery. Counterspeculation, auctions and competitive sealed tenders. *J. of Finance*, 16:8–37, 1961.