Towards Universal Power Efficient Scheduling in Gaussian Channels

Dinesh Rajan, *Member, IEEE*

***Abstract*— In this paper, we propose a framework for designing power efficient schedulers for transmitting bursty traffic sources over Gaussian wireless channels that provides deterministic and statistical guarantees on absolute delays experienced by the source packets. The proposed schedulers compute the transmission rate and power using temporal water-filling techniques without any knowledge of the arrival traffic statistics. The schedulers reduce the average transmission power substantially (55% in some scenarios) for small increases in delay. The framework allows us to design schedulers that artfully tradeoff the performance with the complexity of computing the schedulers. We also introduce an iterative process to compute a lower bound on the transmit power of *any* scheduler that provides absolute delay guarantees. The utility of having accurate traffic predictors is demonstrated; specifically, we show that a perfect one step predictor achieves near optimal performances for small delay bounds. The proposed schedulers and iterative method of computing the lower bound are also shown to provide statistical guarantees on packet delays.**

***Index Terms*— Scheduling, Power control, Data Traffic, Delay bounds.**

I. Introduction

PROVISIONING of quality of service (QoS) is critical for the success of high data rate multimedia services in future wireless networks. In this paper, we propose power efficient schedulers for transmitting bursty traffic sources through a wireless channel. The proposed schedulers provide deterministic or statistical guarantees on absolute packet delays without any prior knowledge of the arrival traffic statistics.

Scheduling is commonly used to refer to the allocation of a particular resource (like bandwidth) to multiple competing entities (like flows) under certain constraints (like fairness); for example, a first come first serve (FCFS) scheduler. In this paper, we consider a single user system with varying number of packet arrivals in every time-slot; such a source is referred to as a bursty source. A scheduler is defined (in Section II) as a mapping from the number of packet arrivals and *state* of the system to the number of packets transmitted. Schedulers are proposed for both stationary and nonstationary traffic sources.

The contributions of this paper are as follows:

- Computes a universal lower bound on the transmit power required by any scheduler that provides deterministic or statistical guarantees on absolute packet delays. The lower bound is computed iteratively and in an off-line manner for a given sequence of packets arrivals. The scheduler which achieves this lower bound is referred to as noncausal minimal power (NOMP) scheduler.
- Proposes a simple memoryless (and causal) scheduler that uses no information on prior scheduled packets or future arrivals to schedule packets: Its performance serves as an upper bound on transmission power and can be computed in closed form. An approximate relationship between transmission power and delay bound is derived. Remarkably, it is shown that the derived approximate relationship is analogous to Shannon's capacity formulation in which the rate is replaced by the effective bandwidth of the source.
- For nonstationary sources, a low complexity water-filling based causal look ahead (CLAD)-0^{th} order[1] scheduler is proposed that has low power consumption.
- For stationary sources, a CLAD-1 scheduler is proposed that uses information of prior arrivals to predict the statistics of future arrivals and schedule packets. The power of this scheduler is lower than that of CLAD-0 schedulers and is near optimal.
- A simple noncausal scheduler, labeled NCLAD-1, is proposed that assumes perfect knowledge of exactly one future arrival. The performance of a NCLAD-1 scheduler shows that an efficient 1-step predictor achieves near optimal performance for all traffic scenarios. Thus, various traffic prediction methods (e.g. neural network based) can be used to construct optimal schedulers.

All the proposed schedulers are solutions of minimizing the transmit power under various constraints: The solutions of these optimization problems (except for CLAD-1 scheduler) can be interpreted as *constrained temporal water-filling* and involves low computational complexity. These schedulers can thus be readily adapted for implementing in various networks including cellular data networks, WLANs, ad-hoc networks and mesh networks. The proposed delay bounded schedulers are conceptually similar to rate and power adaptation in fading channels with short term power constraints [2]. Since the proposed CLAD scheduler achieves performances near that of the NOMP scheduler without any knowledge of arrival traffic, it can be considered to be a preliminary or first generation universal scheduler.

The results in this paper suggest three main messages: i) Irrespective of traffic, simple schedulers exist which are near optimal, ii) Perfect one step traffic predictor achieves near optimal performance for low to medium values of delay, and iii) An approximate formula can be derived for power-delay relationship that depends only on first and second order statistics of the arrival traffic.

Manuscript received May 2006, revised November 2006. This work has been supported by the National Science Foundation under grant CCF-0546519. Part of this paper was presented at the International Conference on Communications (ICC), Paris, June 2004.

The author is with the Department of Electrical Engineering, Southern Methodist University, Dallas, Texas, USA (e-mail: rajand@engr.smu.edu).

Digital Object Identifier 10.1109/JSAC.2007.070516.

[1]The order of the scheduler is defined in Section III.

Recognizing that a substantial portion of the energy consumption in a wireless communication system is at the transmitter, in this paper, we focus on minimizing the transmit power under delay constraints. Power minimization for uplink transmission is important in mobile terminals due to the limited battery resources. Using low power transmission to obtain the desired downlink performance is also important at the base station since it results in reduced interference to other users and consequently increased system throughput.

The importance of incorporating traffic models in the design of wireless communication systems has been well recognized [3], [4]. Techniques that delay data transmission based on channel conditions is commonly used to save transmission power. The transmission scheme which maximizes long-term throughput transmits more power and information in good channel states, and less in poor conditions [5]. Similar concepts are used in the INFOSTATION [6] architectures, where mobile nodes transmit data only when they are close to base-stations; thus, reducing transmission power for increased delays. Schedulers that minimize the transmit power under average delay constraints have been considered in [7]–[10]. These approaches typically use dynamic programming methods to obtain optimal transmission policies. Schedulers that provide QoS guarantees have been an area of active research (see for example [11]); however, many of these have been proposed for constant rate data links like in wired networks. In wireless channels, the instantaneous throughput depends on the channel conditions and the transmit power and is therefore not a constant. There is extensive work on scheduling over wireless channels: see [12]–[16] for a partial list of relevant work. The primary motive behind these works is to efficiently use system resources, often with an aim of fair division of resources. Moreover, a lot of emphasis is given on exploiting the multiuser diversity effects [17], [18].

The remainder of this paper is organized as follows. We introduce some basic notation and formalize the scheduling problem in Section II. Schedulers with deterministic and statistical guarantees on delay are introduced in Sections III and IV respectively. Finally, we conclude in Section V.

II. Problem Setup

Consider a system in which a_n packets arrive at the transmitter at the beginning of time-slot n. Let D_0 be the desired absolute packet delay bound. Delay is measured in terms of the number of time-slots and we use the convention that if packets arriving in time-slot n are transmitted in the same time-slot, then the delay equals 1 time-slot. To deterministically meet the delay bound all a_n packets have to be transmitted within time-slots $n, n+1, \ldots, n+D_0-1$. All arriving packets are stored in a buffer which is assumed to be large enough not to cause any overflows.[2] In this paper, we treat each packet as being infinitely divisible and partial packet transmissions in a time-slot are allowed. For simplicity, we only consider an additive white Gaussian noise (AWGN) channel. The analysis can be easily extended to block fading channels.

[2]If the maximum packet arrival M in any slot is known, then the buffer size needed to prevent overflows is MD_0.

The received signal, Y_n, is given by $Y_n = X_n + z_n$, where X_n is the transmitted signal and z_n is the additive Gaussian noise with variance σ^2. The signals Y_n, X_n and z_n are T_c dimensional vectors, where T_c is the number of symbols in each time-slot. The transmit signal, X_n, depends on the number of packets, u_n, transmitted in time-slot n and the coding and modulation scheme used. Transmit power P_n is chosen to ensure that X_n can be reliably determined from Y_n. In this paper, we consider reliability in the Shannon theoretic sense and use the well known Gaussian capacity formulation [19] to compute the power required to transmit u_n packets as $P_n = P(u_n) = \sigma^2(e^{Ru_n} - 1)$, where R is the size of the packet in bits. Although, such reliability metrics are valid only asymptotically, the performance of practical advanced coding schemes is very close to information theoretic limits. Moreover, similar functional forms like $P(u_n)$ with additional penalty term can be used to model practical systems with finite probability of error. The average power of any scheduler can then be computed as $\frac{1}{N}\sum_{i=1}^{N} P(u_i)$ for a given packet arrival sequence $\{a_n\}_1^N$. Denote by $v_{n,i}$ the number of packets transmitted during time-slot n that arrived during time-slot $n-i$.

Definition 1: A scheduler is defined as a mapping from the number of packet arrivals, a_n and scheduler state, S_n to the number of packets transmitted in different time-slots, $v_{n+i,i}, \forall i \geq 0$, *i.e.*, $(a_n, S_n) \mapsto v_{n+i,i}$.

In other words, at each time-slot n the scheduler computes when each of the a_n packets should be transmitted. Recognize that schedulers which determine the transmission rate (*i.e.*, number of packets transmitted in each time-slot) based only on the total number of packets in the buffer cannot deterministically guarantee the desired delay bound. Thus, buffer state is not a sufficient statistic for guaranteeing a delay bound. The state of the scheduler, S_n, at time-slot n depends on the type of scheduler. For a NOMP scheduler, the state is $S_n = \phi(\{a_i\}_{i=1}^N)$, where N is the length of the arrival sequence and $\phi()$ is a function that represents the relation between packet arrivals and current scheduler state. Although, we do not explicitly characterize this functional form, its value will be clarified in the different cases. For the CLAD schedulers, $S_n = \phi\left(\{a_i\}_{i=1}^{n-1}\right)$ and for the memoryless scheduler $S_n = \Lambda$, the null state. All the proposed schedulers are assumed to be stationary (*i.e.* the mapping is time-invariant) even if the arrivals are not stationary.

For a scheduler that provides deterministic delay guarantees all packets have to be transmitted within D_0 time-slots of arriving at the transmit buffer, *i.e.*, $\sum_{j=0}^{D_0-1} v_{n+j,j} = a_n$. For a scheduler that provides statistical delay guarantees, at most $\delta\%$ of the packets may violate the delay bound, *i.e.*,

$$Pr\{d_i > D_0\} \leq \delta, \tag{1}$$

where d_i is the absolute delay of the i^{th} packet and δ is the fraction of packets that violate the delay bound. In this paper, we assume that packets which violate the delay bound

are dropped.[3] Thus, $v_{n,i} = 0$ for $i \notin \{0, 1, \ldots, D_0 - 1\}$. The number of packets, u_n, transmitted in time-slot n is given by $u_n = \sum_{j=0}^{D_0-1} v_{n,j}$. In the next section, we compute power efficient schedulers that provide deterministic delay guarantees.

Note that with no constraint on the maximum packet arrivals within a time-slot, an absolute delay bound cannot be guaranteed with finite transmission power. In power constrained systems, an outage formulation is more applicable [20]. We do not consider power limited systems in this paper; our goal is to characterize bounds on power required to provide certain delay guarantees to the traffic. As one application of our results, we show in Section IV how increasing the ratio of packets that violate delay bound can be used to provide service in power constrained systems.

III. Schedulers with Deterministic Delay Guarantees

In this section, we compute power optimal schedulers with different amounts of information available to the scheduler. We first compute a noncausal scheduler, whose performance serves as a lower bound on the performance of any scheduler and then proceed to compute different causal schedulers that approach the performance lower bound.

A. Noncausal minimal power (NOMP) scheduler

We first compute a noncausal scheduler that uses information about the entire arrival sequence (*i.e.*, $S_n = \phi(\{a_i\}_{i=1}^N)$) to schedule packets and minimize the average power consumption. The problem of interest can be formally stated as follows,

$$P^*_{NOMP} = \min_{\substack{\{v_{i,j}\} \\ i=1\ldots,N,\ j=0,\ldots,D_0-1}} \frac{1}{N}\sum_{i=1}^{N} P\left(\sum_{j=0}^{D_0-1} v_{i,j}\right)$$
$$0 \le v_{i,j};\ \sum_{j=0}^{D_0-1} v_{i+j,j} = a_i, i = 1, \ldots, N \quad (2)$$

It should be noted that a boundary condition on $v_{k,j}$ imposed due to causality is not explicitly mentioned in (2). Specifically, recognize that $v_{k,j} = 0$, for $k < D_0$ and $j \ge k$, since there are no packet arrivals before time-slot 0. In this case, the scheduler state S_n can be considered as a D_0 length vector with elements,

$$\left[\sum_{\substack{j=0 \\ j\neq 0}}^{D_0-1} v_{n,j},\ \sum_{\substack{j=0 \\ j\neq 1}}^{D_0-1} v_{n+1,j}, \ldots,\ \sum_{\substack{j=0 \\ j\neq D_0-1}}^{D_0-1} v_{n+D_0-1,j}\right].$$

Recall that the scheduler computes the number of packets to transmit and the power based on S_n and a_n.

For any given sequence of packet arrivals $\{a_n\}_1^N$, it can be shown that (2) is a convex optimization problem and has a minimum. In Appendix A, we give an iterative method for computing P^*_{NOMP}. The iterative procedure schedules packet transmissions using information of future arrivals in a noncausal manner. Hence, the scheduler is referred to as a noncausal minimal power (NOMP) scheduler and its performance serves as a lower bound on the power of *any* scheduler that guarantees delay bound D_0 for that arrival sequence. Numerical values of the lower bound are given in Figure 4 and explained in Section III-G.

B. Causal look ahead (CLAD)-0 schedulers

We now propose a series of causal schedulers labeled CLAD-k schedulers in which the order of the scheduler, k, is the number of future time-slots up to which arrivals are predicted. The rationale is that for stationary sources, the statistics of future arrivals can be estimated from prior arrivals. The CLAD-0 scheduler does not predict the statistics of future arrivals and can be used even when the traffic is nonstationary.

To derive the CLAD-0 scheduler, we modify (2) to minimize the local average power rather than global average power. The state of the scheduler $S_n = \phi(\{a_i\}_{i=1}^n)$. In this case, S_n can be considered as a D_0 state vector with elements $\left[\sum_{j=1}^{D_0-1} v_{n,j},\ \sum_{j=2}^{D_0-1} v_{n+1,j}, \ldots,\ \sum_{j=D_0-1}^{D_0-1} v_{n+D_0-2,j},\ 0\right]$. At time-slot n, the CLAD-0 scheduler computes $v_{n+j,j}, j = 0, 1, \ldots, D_0 - 1$ to minimize $\sum_{j=0}^{D_0-1} P(\tilde{u}_{n+j})$, where $\tilde{u}_{n+j} = \sum_{k=j}^{D_0-1} v_{n+j,k}$ is the number of packets scheduled for transmission at time-slot $n+j$ that arrived during or before time-slot n. Note that this scheduler uses the knowledge of prior scheduled transmissions, $v_{n+j,i}, i = j+1, \ldots, D_0$ via S_n to compute $v_{n+j,j}$. Formally,

$$P^*_{CLAD-0} = \min_{\substack{\{v_{n+j,j}\} \\ j=0,\ldots,D_0-1}} \frac{1}{D_0}\sum_{i=0}^{D_0-1} P\left(\sum_{k=i}^{D_0-1} v_{n+i,k}\right)$$
$$0 \le v_{n+j,j},\ \sum_{j=0}^{D_0-1} v_{n+j,j} = a_n \quad (3)$$

It can be easily shown that the solution to (3) is given by temporal water-filling as

$$v_{n+j,j} = \left(\alpha - \sum_{k=j+1}^{D_0-1} v_{n+j,k}\right)^+,\ j = 0, \ldots, D_0 - 1 \quad (4)$$

where $(x)^+ = \max\{x, 0\}$, and α is computed from $\sum_{j=0}^{D_0-1} (\alpha - v_{n+j,j})^+ = a_n$. The performance of this scheduler is shown in Figure 4 and is discussed in Section III-G.

C. CLAD-1 scheduler

For stationary sources, we propose a scheduler which minimizes local transmit power by estimating the distribution of future arrivals. In this paper, we use a simple histogram to

[3] An alternative formulation is to ensure that all packets have to be completely transmitted, but up to $\delta\%$ of them may violate the delay bound. In this case, packets which violate the delay bound are queued separately (rather than dropped) and transmitted when $u_n \le \mathbb{E}[a_n]$, *i.e.*, when there are not too many packets scheduled for transmission.

estimate the distribution of future arrivals and thus $\hat{p}(a_{n+1} = i) = \frac{1}{n}\sum_{k=1}^{n} I(a_k = i), \forall i$, where $I()$ is the indicator function. As n increases $\hat{p} \rightarrow p$, the true distribution of the source. At each time-slot n the CLAD-1 scheduler not only computes $v_{n+j,j}$ but also $v^{(k)}_{n+j+1,j}$ for $k = 1, \ldots, M_n = \max\{a_1, \ldots, a_n\}$. Note that $v^{(k)}_{n+j+1,j}$ represents the number of packets transmitted in time-slot $n+j+1$ if k packets arrive at time-slot $n+1$. Formally, the scheduler is chosen to optimize

$$P^*_{CLAD-1} = \min_{\substack{\{v_{n+j,j}, v^{(k)}_{n+j+1,j}\} \\ j=0,1,\ldots,D_0-1}} \sum_{k=0}^{M_n}\sum_{i=0}^{D_0} P(\hat{u}_{n+i})p(a_{n+1}=k)$$
$$0 \le v_{n+j,j}, \sum_{j=0}^{D_0-1} v_{n+j,j} = a_n, \quad (5)$$
$$\sum_{j=0}^{D_0-1} v^{(k)}_{n+j+1,j} = k \ \forall 1 \le k \le \max\{a_1, \ldots, a_n\}$$

where $\hat{u}_{n+i} = \sum_{j=0}^{D_0-1} v_{n+i,j}$. Solving (5) analytically is intractable and hence we resort to numerical optimization techniques. The performance of the CLAD-1 scheduler is given in Figure 4 and explained in Section III-G.

The CLAD-m schedulers can be derived using similar techniques: However, their computational complexity increases with m. Moreover, as will become evident from Section III-G the performance of the CLAD-1 scheduler is close to the NOMP scheduler and thus is a reasonable choice for obtaining good performance at low complexity. It should be reiterated that the CLAD-m schedulers assume stationarity of the arrival traffic and are not guaranteed to provide significant scheduling gains when the traffic is non-stationary. The formulation of the CLAD-m schedulers are not unique. For example, a lower complexity CLAD-1 scheduler formulation assumes that when computing the scheduler at time-slot n, the a_{n+1} packets are uniformly split and transmitted over D_0 time-slots. A similar approach is taken in the NCLAD-m schedulers which is discussed next.

D. NCLAD-1 scheduler

The main motivation behind the NCLAD-1 scheduler is to understand and quantify the gains in using traffic predictors in the scheduling context. The NCLAD-1 scheduler minimizes local transmit power assuming that one future arrival is known exactly, *i.e.*, at time-instant n, scheduler has knowledge of $\{a_i, i = 1, \ldots, n+1\}$. Thus, scheduler state $S_n = \phi(\{a_i\}_1^{n+1}) = \left[\sum_{\substack{j=0 \\ j\neq 0}}^{D_0-1} v_{n,j}, \sum_{\substack{j=0 \\ j\neq 1}}^{D_0-1} v_{n+1,j}, \sum_{\substack{j=1 \\ j\neq 2}}^{D_0-1} v_{n+2,j}, \ldots, \sum_{\substack{j=D_0-2 \\ j\neq D_0-1}}^{D_0-1} v_{n+D_0-1,j}\right]$; note the noncausal nature of S_n which depends on a_{n+1}. Moreover, the NCLAD-1 scheduler assumes that the a_{n+1} packets are uniformly divided and transmitted over time-slots $n+1, \ldots, n+D_0$. [4]

Similar to the earlier cases, the optimization problem is posed as,

$$P^*_{NCLAD-1} = \min_{\substack{\{v_{n+j,j}\} \\ j=0,\ldots,D_0-1}} \frac{1}{D_0}\sum_{i=0}^{D_0-1} P\left(\sum_{k=(i-1)^+}^{D_0-1} v_{n+i,k}\right).$$
$$0 \le v_{n+j,j}\ , \sum_{j=0}^{D_0-1} v_{n+j,j} = a_n \quad (6)$$

The solution is again obtained by temporal water-filling as

$$v_{n+j,j} = \begin{cases} \left(\alpha - \sum_{k=j+1}^{D_0-1} v_{n+j,k}\right)^+, & j = 0 \\ \left(\alpha - \sum_{k=j+1}^{D_0-1} v_{n+j,k} - \frac{a_{n+1}}{D_0}\right)^+, & 1 \le j < D_0 \end{cases} \quad (7)$$

where α is computed as $\sum_{j=0}^{D_0-1} (\alpha - v_{n+j,j})^+ = a_n$. Although, there are many different schedulers that use knowledge of future arrivals, our results indicate that the proposed NCLAD-1 scheduler is near optimal for small delay bounds. Moreover, the complexity of the proposed NCLAD-1 scheduler is low. Numerical values of the NCLAD-1 scheduler is given in Figure 6 and explained in Section III-G. As noted earlier, the main objective of the NCLAD-1 scheduler is to study the effect of perfect one step traffic predictor on scheduler performance.

E. Memoryless scheduler

We now present a simple memoryless scheduler in which $v_{n+j,j}, j = 0, 1, \ldots, D_0 - 1$ depends only on a_n and not on the prior scheduled packets; thus $S_n = \Lambda$, the Null state. Specifically, the memoryless scheduler uniformly spreads the packet transmissions across the entire available time-slots, *i.e.*, $v_{n+j,j} = \frac{1}{D_0} a_n$ and thus, $u_n = \frac{1}{D_0}\sum_{j=0}^{D_0-1} a_{n-j}$. This memoryless scheduler serves four objectives: i) Its performance serves as an upper bound on the transmit power of any scheduler for arbitrary delay bounds, ii) Its performance can be derived analytically in closed form, if the traffic is stationary. iii) It is used to show the connection between filtering and scheduling and derive the delay-bandwidth relationship, and iv) An approximate formula for the power of this scheduler at large delay bounds is derived that depends only on the first and second order statistics of the incoming traffic. Remarkably, this approximate formulation is similar to the notion of effective bandwidth [21]. For non-stationary sources, the first and second order statistics in the approximate formula are replaced by the sample mean and variance for that arrival sequence.

[4] An approach similar to the CLAD-1 scheduler can be used in which the scheduler at time n *attempts* to optimize the transmission of the a_{n+1} packets also; however, as the results indicate, the proposed NCLAD-1 scheduler attains near optimal performance for small delays.

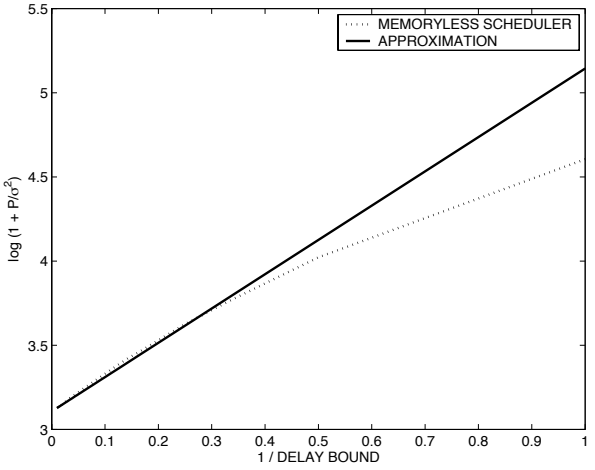


Fig. 1. Performance of approximate power derivation for Ethernet traffic

1) Closed form analysis: The average power of the memoryless scheduler, $P_{memoryless}$, is given by

$$P_{memoryless} = \frac{1}{N}\sum_{n=1}^{N} P\left(\frac{1}{D_0}\sum_{j=0}^{D_0-1} a_{n-j}\right).$$

If the source arrivals are stationary, then $P_{memoryless}$ can be derived in closed form as

$$\begin{aligned} P_{memoryless} &= \sum_{a_1 \ldots a_{D_0}} P\left(\frac{\sum_{i=1}^{D_0} a_i}{D_0}\right) p(a_1, \ldots, a_{D_0}) \quad (8) \\ &= \sigma^2\left(\sum_{a_1 \ldots a_{D_0}} \prod_{i=1}^{D_0} e^{Ra_i/D_0} p(a_i) - 1\right) \\ &= \sigma^2\left(\prod_{i=1}^{D_0} \sum_{a_i} e^{Ra_i/D_0} p(a_i) - 1\right) \\ &= \sigma^2\left(\left(\mathbb{E}[e^{Ra_n/D_0}]\right)^{D_0} - 1\right), \quad (9) \end{aligned}$$

where $p(a_1 \ldots, a_{D_0}) = \prod_{i=1}^{D_0} p(a_i)$ is the joint pmf of the arrivals which is assumed to be an i.i.d. process. Note that the memoryless scheduler is derived without assuming stationarity of the arrival process. The stationarity of the arrival process is only used to calculate the average power in closed form (9).

Rewriting (9), we see that

$$\log\left(1 + \frac{P}{\sigma^2}\right) = D_0 \log \mathbb{E}[e^{Ra_n/D_0}] \quad (10)$$

It can be seen that as $D_0 \to 1$, the RHS equals $\log \mathbb{E}[e^{Ra_n}]$. Also, as $D_0 \to \infty$, by the law of large numbers (applied to (8)), the RHS equals $\mathbb{E}[Ra_n]$ which is the Shannon limit.

2) Performance analysis based on first and second order statistics: To better understand the performance of the memoryless scheduler, we derive an approximate relationship for $P_{memoryless}$ that depends only on the first order (mean) and second order (variance) statistics of the arrival process. We then study the validity of this approximation when applied to traffic for which the entire statistics are unknown or difficult to model, e.g. Ethernet traffic. The approximation derived in Appendix B is given as follows:

$$\log\left(1 + \frac{P}{\sigma^2}\right) = R(\lambda + \frac{1}{2}\frac{R\sigma_a^2}{D_0}) = R(\lambda + \frac{1}{2}R\sigma_u^2), \quad (11)$$

where σ_u^2 is the variance of the output traffic and is derived in Appendix B. The accuracy of the approximation (11) at high delays is evident from Figure 1 which compares the performance of the memoryless scheduler and the approximate formulation for Ethernet traffic. Interestingly, the approximation is accurate only at very high delays (on the order of 100 time-slots or higher) for Ethernet traffic. However, for MPEG and i.i.d traffic, the approximation holds at smaller values of delay bound.[5] Moreover, the analysis easily extends to the case of multiple flows with different delay constraints and our results will be presented at a later forum.

3) Non-monotonic behavior of memoryless scheduler: It should be noted that the power of the memoryless scheduler is *not* a monotonically decreasing function of the delay bound. A simple example illustrates this non-monotonic behavior. Consider the deterministic arrival sequence $\{0, M, 0, M, \ldots\}$. For this arrival sequence and a delay bound of 2 time-slots, the memoryless scheduler transmits $M/2$ packets in every time-slot: Thus, the average power equals $P_{memoryless}(2) = \sigma^2(e^{RM/2} - 1)$. For a delay bound of 3 time-slots, the scheduler alternatingly transmits $M/3$ and $2M/3$ packets and hence the average power equals $P_{memoryless}(3) = \sigma^2((e^{RM/3} + e^{2RM/3})/2 - 1)$. Clearly $P_{memoryless}(3)$ is greater than $P_{memoryless}(2)$ due to the convexity of the exponential function.

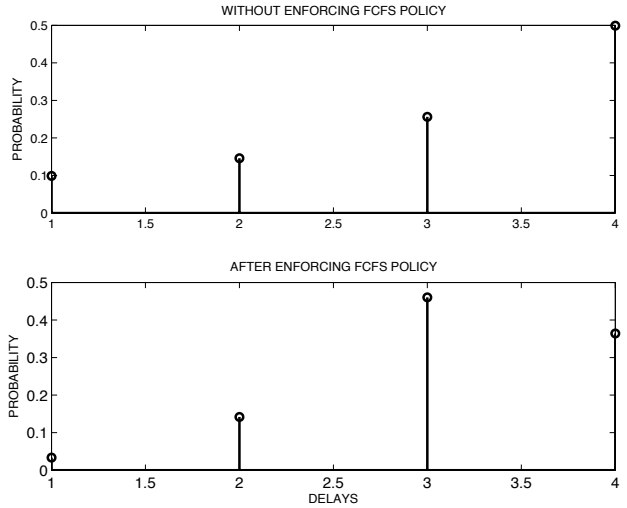
4) Average delay and delay distribution: The average delay of the proposed scheduler can be easily derived from Little's law as the ratio of average buffer length $\mathbb{E}[x_n]$ and average arrival rate. The buffer length x_n for the memoryless scheduler is given by

$$\begin{aligned} x_n &= \sum_{i=1}^{n} a_i - \sum_{i=1}^{n-1} u_i \\ &= \frac{D_0 - 0}{D_0} a_n + \frac{D_0 - 1}{D_0} a_{n-1} + + \ldots + \frac{1}{D_0} a_{n-D_0+1} \\ &= \sum_{i=0}^{D_0-1} \left(1 - \frac{i}{D_0}\right) a_{n-i} \end{aligned}$$

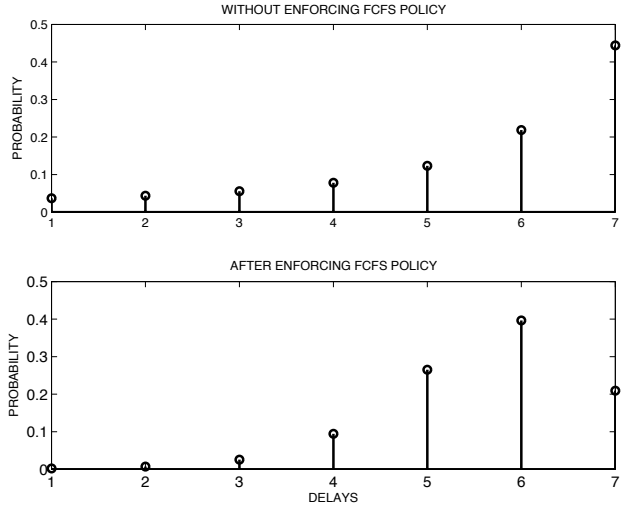
It follows that $\mathbb{E}[x_n] = \mathbb{E}[a_n]\left(\frac{D_0+1}{2}\right)$ and hence the average delay equals $\frac{D_0+1}{2}$. For the other schedulers proposed, deriving the average delay in closed form is not feasible. A trivial upper bound on the average delay for all the proposed schedulers is D_0.

It should be noted that for a specified absolute delay bound, the variance in the delay histogram, *i.e.*, the actual delays experienced by the different packets, can be reduced without any additional increase in transmission power by applying traditional packet re-ordering mechanisms like FCFS policy. The application of the FCFS policy does not change the packet delay distribution when $D_0 = 2$. The delay distribution before and after enforcing FCFS policy is given in Figure 2 for the CLAD-0 scheduler. Clearly, the concatenation of FCFS

[5]Plots similar to Figure 1 for MPEG and i.i.d. traffic are not shown.



(a)



(b)

Fig. 2. Histogram of delay distributions with and without imposing FCFS policy in conjunction with the proposed CLAD-0 scheduler: a) Delay bound of 4 time-slots, b) Delay bound of 7 time-slots.

policy reduces the variance of this distribution at all delays. It should be noted that by coupling with a FCFS policy, no packets are received out of order, a useful feature for real time traffic source and TCP based networks. In case of multiple flows with different deadlines, the proposed schedulers can be easily coupled with other reordering mechanisms like earliest-deadline-first (EDF) [11].

F. Water-filling interpretation of schedulers

The solution of the proposed optimization problems can be interpreted using standard water-filling method. This water-filling process is pictorially depicted in Figure 3. Consider a multiple binned vessel, where each bin corresponds to one time-slot and the base of all bins are of unit area. The volume of water poured equals a_n for the memoryless and CLAD-0 schedulers. The memoryless scheduler, is similar to water-filling into D_o empty bins, *i.e.*, the a_n packets are distributed uniformly across all D_0 bins. The CLAD-0 scheduler is

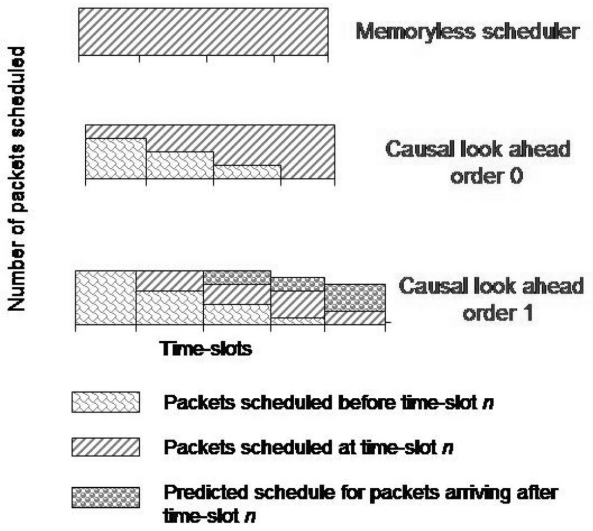


Fig. 3. Water-filling interpretation of schedulers

analogous to pouring the water into a vessel with D_0 bins; the vertical level in the first $D_0 - 1$ bins is proportional to the number of packets already scheduled for transmission ($v_{n+i,j}$) and the last bin is empty.

The NOMP scheduler is similar to water-filling into a bin with $N + D_0 - 1$ bins. The amount of water poured equals $\sum_{n=1}^{N} a_n$. However, there is a *valve* or *filter* between the bins that allows only certain fluids to pass through in both directions. The different fluids will settle down to as uniform a level as possible in each bin subject to valve operation constraints.

G. Numerical Results

We now numerically study the performance (average power versus delay bound) of the different schedulers.

Scheduler Performance for i.i.d. traffic: The average power of the different schedulers are plotted against the delay bound in Figure 4 for an i.i.d. arrival sequence of length 10,000 time-slots. Clearly, at a delay of 1 time-slot, all packets have to be transmitted in the same time-slot that they arrive. Thus, all schedulers require the same power. As the delay bound increases from 1 to 2 time-slots, the power required decreases substantially (nearly 55% for CLAD-0 scheduler) for all schedulers. Eventually, as the delay goes to infinity, the required power approaches the Shannon limit. The Shannon limit is simply given by $P\left(\frac{R}{N}\sum_{n=1}^{N} a_n\right)$ and is the power required to transmit constant rate traffic.

The power required by the memoryless scheduler is higher than the power required by the CLAD-0 scheduler, which is not surprising since the memoryless scheduler does not utilize knowledge of prior packet schedules in determining the transmission rate. As the delay increases the CLAD-0 scheduler smoothes the input traffic thereby reducing *burstiness* of output traffic, an effect which has also been observed in case of schedulers which minimize power under average delay constraints [8]. The CLAD-1 scheduler has performance

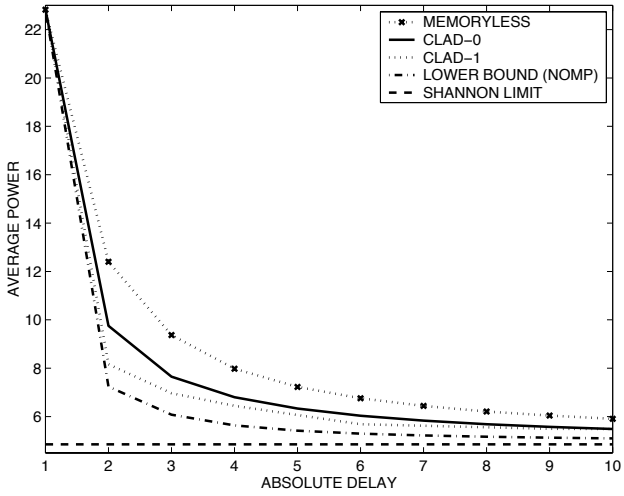


Fig. 4. Average power versus absolute delay for the proposed schedulers.

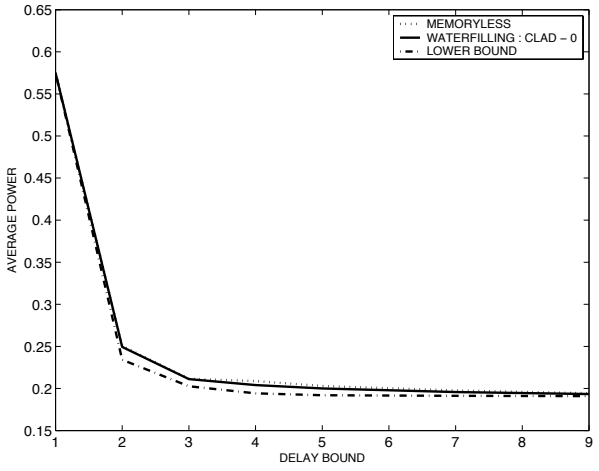


Fig. 5. Performance of proposed schedulers for MPEG traffic.

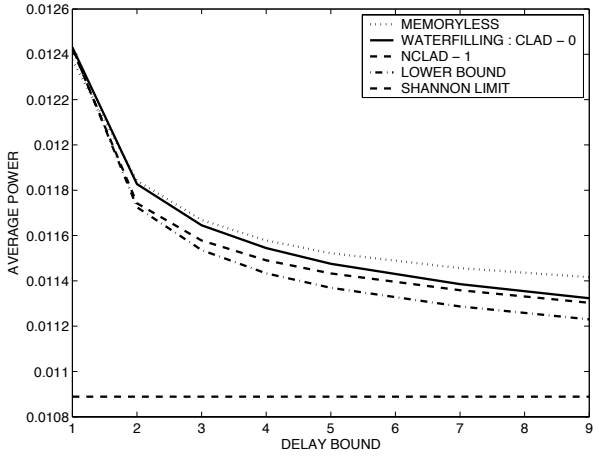


Fig. 6. Performance of proposed schedulers for Ethernet traffic.

better than even the CLAD-0 scheduler and nearly that of the NOMP scheduler. It should be reiterated that only the CLAD-1 scheduler requires the stationarity of arrival process. The other schedulers can be used even for nonstationary traffic sources. In this case, the performance of the NCLAD-1 scheduler was similar to that of the CLAD-1 scheduler and is not shown.

Scheduler Performance for MPEG traffic: The performance of the proposed schedulers is given in Figure 5 for an MPEG traffic source. The average power of the memoryless, CLAD-0 and NOMP schedulers are plotted in Figure 5. It can be seen that the average power decreases substantially (nearly 60%) for small increase in delay. Further, the performance of all three schedulers are nearly identical. Thus, for this traffic source, the memoryless and CLAD-0 schedulers exhibit near optimal performance. Analysis of the actual traffic sequence illuminates the reasons for such scheduler behavior. The MPEG traffic consists of packets that contain periodic intracoded frames (I-frames) with intercoded frames (P/B frames) in between. The I-frames contain significantly more data than the P/B frames and thus the power required to transmit the I-frames is significantly higher than the power required to transmit the P/B-frames. Hence, changing the transmission rates of the P/B frames according to prior I-frame packets scheduled and vice-versa does not provide significant power reduction over the memoryless scheduler.

Scheduler Performance for wide-area traffic: The performance of the proposed schedulers for Ethernet traffic is given in Figure 6. The trace files for the simulations were obtained from [22] and details of the trace files are given in [23]. In this case, the rate of decrease of power with delay is slower than for i.i.d. or MPEG traffic. The reason for the slower decrease in average power versus delay bound curve is the self similar nature of the traffic stream, *i.e.*, the traffic exhibits burstiness across an extremely wide range of time-scales [24]. Thus, smoothing the traffic at smaller time-scales (smaller delay bounds) does not completely remove the burstiness at larger time-scales and hence the power of the CLAD-0 and memoryless schedulers are significantly higher than the NOMP scheduler.

Also, the NOMP scheduler has significantly better performance than the CLAD-0 scheduler. However, for this arrival traffic, as expected, the CLAD-1 scheduler does not provide significant gain over the CLAD-0 scheduler since the arrival traffic is not stationary and a simple histogram based predictor is not very accurate. The performance of the NCLAD-1 scheduler is near optimal for small delays. Thus, illustrating that with a perfect one-step predictor near optimal power is obtained for small delays. We found that the NCLAD-2 scheduler, which has perfect knowledge of 2 future arrivals, nearly achieves NOMP scheduler performance for even higher delays than the NCLAD-1 scheduler. For clarity the performance of the NCLAD-2 scheduler is not shown in Figure 6.

IV. Schedulers with Statistical Delay Guarantees

In this section, we modify the scheduler design to provide statistical delay guarantees on the traffic.

A. Statistical NOMP scheduler

For a given arrival sequence $\{a_n\}_1^N$ we now compute a lower bound on the power required by any scheduler which provides delay guarantees of the form (1). In this paper, we assume that packets which violate the delay bound are dropped

and retransmission will be ensured by higher layers.[6] The statistical NOMP scheduler is computed in two steps:

1) *Deterministic NOMP:* Calculate $v_{i,j}$ and u_n like in the deterministic NOMP scheduler (Section III-A).

2) *Dropping policy:* Given δ and $\{a_n\}_1^N$, compute a dropping threshold u_{dr} such that no more than u_{dr} packets are transmitted in any time-slot n. At each time-slot n the remaining $u_n - u_{dr}$ packets are dropped. The threshold u_{dr} is determined to ensure that the total fraction of dropped packets equals δ *i.e.*, $\sum_{n=1}^{N}(u_n - u_{dr})^+ = \delta \sum_{n=1}^{N} a_n$. The dropping policy is essentially an inverse water-filling process, *i.e.*, a threshold is set and all packets larger than the threshold are dropped. The optimality of this NOMP scheduler is discussed in Appendix D.

The performance of the statistical NOMP scheduler is given in Figure 7a for two different values of δ. For a given value of δ, the average power is a decreasing function of the delay bound similar to the earlier case of absolute delay bounds. This rate of decrease is smaller for larger values of δ. The reduction in power with increasing δ is clear for all delays from the figure.

The proposed statistical NOMP scheduler is constructed to ensure that over the entire arrival sequence $\{a_i\}_1^N$ not more than $\delta\%$ of the packets violate the delay bound. However, this scheduler does not guarantee that over all subsets $\{a_i\}_1^n$ lesser than δ fraction of packets will violate the delay bound.

B. Statistical CLAD schedulers

The proposed deterministic CLAD schedulers can be readily modified to guarantee statistical bounds on delay (1). There are three main steps in the calculation of the statistical CLAD schedulers namely:

1) *Deterministic CLAD:* Compute $v_{k,j}$ and u_n as given in Section III-B.

2) *Adaptive dropping policy:* At each time-slot n, compute $\eta_{al}(n) = \delta \sum_{i=1}^{n} a_n$ the total number of packets that could have violated the delay bound and $\eta_{act}(n)$ the actual number of dropped packets. Packets are dropped if $u_n > u_n^{th}$ and $\eta_{act}(n) \leq \eta_{al}(n)$; the number of dropped packets equals $u_n - u_n^{th}$, and u_n^{th} is the dropping threshold. The rationale behind this dropping policy comes from the realization that the transmission of large number of packets consume exponentially high power. Hence, the maximum packet transmission size is limited by setting a adaptively varying threshold.

3) *Threshold update:* The dropping threshold is adapted as follows: $u_{n+1}^{th} = u_n^{th} + \Delta_{th}(\eta_{act}(n) - \eta_{al}(n))$, where Δ_{th} is the threshold updating step parameter.

C. Numerical Results

The performance of the statistical CLAD-0 scheduler is given in Figure 7a for two different values of δ. For large values of δ the reduction in power with increasing delays is negligible, since most packets are dropped and there is not much variation in the transmission rate. As $\delta \to 1$, the required power reaches zero for all delay bounds since all packets are dropped. As $\delta \to 0$, the required power is the same as that of the deterministic schedulers. The performance of statistical CLAD-1 scheduler is also given in Figure 7a. It can be clearly seen from the figure that the performance of the CLAD-1 scheduler is better than the CLAD-0 scheduler but not as good as the CLAD-1 scheduler. In contrast to the statistical NOMP scheduler, the statistical CLAD schedulers are designed to ensure that over all finite time intervals $\{a_i\}_{i=1}^n$ no more than $\delta\%$ of the packets violate the delay bound. The variation of power of the CLAD-0 scheduler with δ is given in Figure 7b for two different delay bounds. As expected the power decreases rapidly with δ and eventually approaches 0 as $delta \to 1$. Not surprisingly, the rate of decrease is higher for smaller delay bounds, since in that case the output traffic is more bursty than at higher delay bounds.

D. Approximate analytical bound: Statistical scheduler

Proceeding as before, we derive in Appendix C the following simple approximation for the behavior of statistical scheduler:

$$\log\left(1-\delta+\frac{P}{\sigma^2}\right) = R\left(\lambda+\frac{R\sigma_u^2}{2}\right) + \log\left(\Phi\left[\Phi^{-1}(1-\delta) - R\sigma_u\right]\right), \quad (12)$$

where $\Phi(.)$ is the standard Normal CDF. The numerical accuracy of the approximation is evident from Figure 7a. It should be noted that this closed form approximation is mainly used to observe trends in variation of power with delay, average traffic rate and packet loss probabilities.

As one application of the approximation, in power constrained systems, one could drop larger fraction of packets to satisfy power constraints. For large SNR (P/σ^2), we can compute this fraction of packets to drop for a given power P_0 and traffic from (12) as

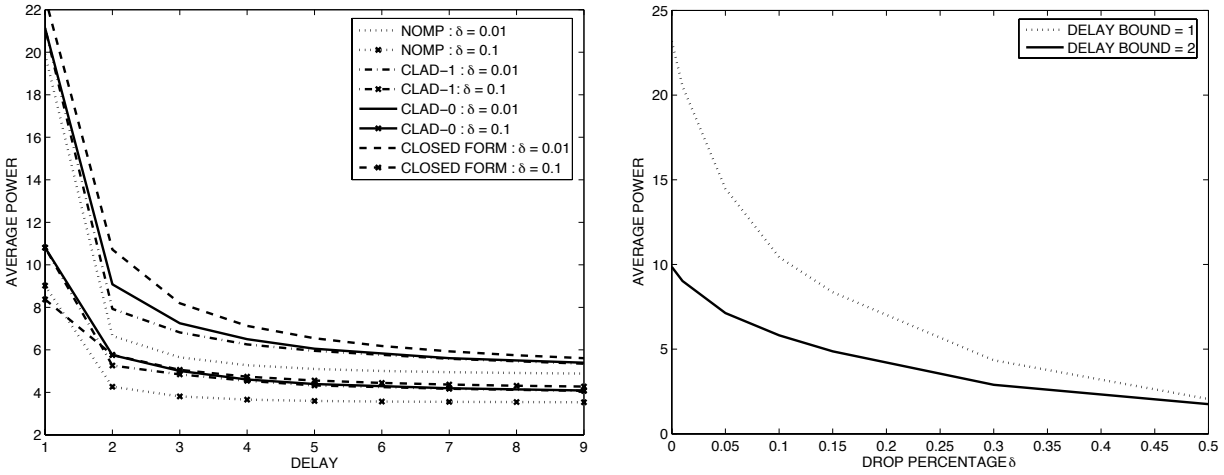
$$\delta = 1 - \Phi\left(R^2\sigma_u + \Phi^{-1}\left(\frac{(1+\frac{P_0}{\sigma^2})}{e^{R\lambda+R^2\sigma_u^2/2}}\right)\right) \quad (13)$$

There are many variations of the adaptive dropping and thresholding methods. Investigating other adaptation policies that achieve performance closer to the statistical NOMP scheduler should be considered in future work.

V. Conclusions

In this paper, we introduced power efficient schedulers which provide deterministic and statistical guarantees on packet delays. The proposed schedulers achieve near optimal performance without prior knowledge of arrival traffic statistics. A universal lower bound is proposed that provides a lower bound on the performance of any scheduler that guarantees a desired delay bound. We believe that the lower bound introduced is not a tight bound for the class of causal schedulers and obtaining tighter bounds should be considered in future work. The proposed schedulers can be easily extended to include multiple flows with different delay constraints. For

[6]Moreover, we assume partial packets can be dropped and retransmitted. Packet integrity constraints can be easily imposed in this framework by considering the optimal continuous valued solution and mapping them to discrete values.



(a) (b)

Fig. 7. a) Performance of proposed statistical NOMP, CLAD-1 and CLAD-0 schedulers as a function of delays for two different values of δ. The closed form approximation is also shown in the Figure. b) Variation of the power of CLAD-0 scheduler with δ for two different values of delays: The performance of statistical NOMP and CLAD-1 schedulers are similar and not shown.

multiple accessing schemes that are orthogonal in frequency or code space, the scheduling concept can be directly applied with appropriate modification to the power-rate formula. For multiple accessing schemes, based on time division multiplexing the proposed formulation can be applied using a modified form of the delay calculation. Extension to multi-hop scenarios should be considered in future research.

ACKNOWLEDGMENTS

The author thanks A.Sabharwal for some enlightening discussions on the paper contents.

APPENDIX

A. Iterative water-filling

First, it should be noted that optimization problem (2) does not have a unique minimizer. The objective function in (2) can be shown to be a convex function and the constraint set can be shown to be a convex set; thus, the local optimum equals the global optimum. By a simple change of variables, optimization problem (2) can be rewritten as

$$P^*_{NOMP} = \min_{\{u_i\}} \frac{1}{N}\sum_{n=1}^{N} P(u_n) \quad (14)$$
$$0 \le u_n, \quad \sum_{j=1}^{n-D_0+1} a_j \le \sum_{j=1}^{n} u_j \le \sum_{j=1}^{n} a_j$$

The Hessian matrix for the objective function (14) can be computed and shown to be positive definite and thus (14) is a strictly convex function. Thus, (14) has a unique minimizer. The equivalence between the two optimization problems (2) and (14) can be explained based on the relationship between $\{v_{i,j}\}$ and $\{u_i\}$. For any given set $\{v_{i,j}\}$ that satisfies the constraints to (2), we can compute a set $\{u_i\}$ that satisfies constraints to (14) by choosing $u_i = \sum_{j=0}^{D_0-1} v_{i,j}$. Similarly, for every set of $\{u_i\}$ that satisfies constraints to (14) we can compute at least one set of $\{v_{i,j}\}$ that satisfies constraints to (2). For example, when $D_0 = 2$, set $v_{1,0} = u_1, v_{2,1} = a_1 - u_1, v_{2,0} = u_2 - v_{2,1}, \ldots$. In general, more than one set $\{v_{i,j}\}$ corresponds to a given set $\{u_i\}$. Hence, the optimization problems (2) and (14) have the same minimum.

Now, we explain the iterative process used to solve (2). At the k^{th} iteration denote by $v^k_{n,i}$ the number of packets transmitted at time slot n that arrived at time $n-i$. Consider an arrival sequence $\{a_n\}$ of length N. At each time-slot n compute $v^k_{n+i,i}$ for $i = 0, 1, \ldots, D_0 - 1$ based on two factors: i) The number of packets scheduled for transmission in time-slots $n+1, ..., n+D_0-1$ by the $(k-1)^{th}$ iteration among the packets that arrived during time-slots $n+1, \ldots, n+D_0-1$, and ii) The number of packets scheduled for transmission in time-slots $n, n+1, \ldots, n+D_0-1$ by the k^{th} iteration among the packets that arrived during time-slots $a_{n-1}, \ldots, a_{n-D_0}$. Further, $v^k_{n+i,i}$ is computed to minimize $P\left(\sum_{j=0}^{D_0-1} \tilde{u}^k_{n+j}\right)$ where $\tilde{u}^k_{n+j} = \sum_{i=j}^{D_0-1} v^k_{n+j,i} + \sum_{i=0}^{j-1} v^{k-1}_{n+j,i}$. The solution to this optimization problem is given by water-filling techniques as

$$v^k_{n+j,j} = \left(\beta - \sum_{i=j+1}^{D_0-1} v^k_{n+j,i} - \sum_{i=0}^{j-1} v^{k-1}_{n+j,i}\right)^+ \quad (15)$$

where β is computed from $\sum_{j=0}^{D_0-1} \left(\beta - v^k_{n+j,j}\right)^+ = a_n$ The average power after the k^{th} iteration is given by $P^{(k)} = \frac{1}{N}\sum_{i=1}^{N} P\left(u^k_i\right)$, where $u^k_i = \sum_{j=0}^{D_0-1} v^k_{i,j}$. The variation of the power with iteration is given in Figure 8, which shows that the iterative process converges within a few iterations for all delay bounds. As noted earlier, since the problem is convex this iterative procedure converges to the global minimum.

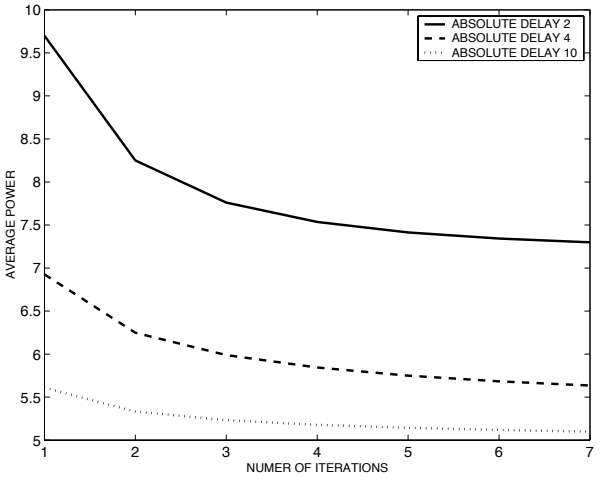


Fig. 8. Convergence of iterative process in computing the lower bound on powers at different absolute delay constraints.

B. Approximate closed form analysis of memoryless scheduler

For i.i.d. traffic arrivals, the distribution of the output packets for delay bound D_0 is given by the convolution of the distribution of the arrival traffic. For example, when $D_0 = 2$, the output distribution is given by

$$\begin{aligned} Pr(u_n = i) &= \sum_{k=0}^{2i} Pr(a_{n-1} = k, a_n = 2i - k) \\ &= \sum_{k=0}^{2i} Pr(a_{n-1} = k) Pr(a_n = 2i - k) \end{aligned}$$

where the joint distribution is written as the product of the marginals due to the *i.i.d.* nature of arrival traffic. A similar expression involving integrals results if the input distribution is continuous.By the Central Limit Theorem, for large delays, the output distribution may be approximated as a Gaussian distribution. The mean of the output distribution equals $\mathbb{E}[a_n]$ and the variance may be computed as $\sigma_u^2 = \mathbb{E}[(\frac{a_1+\ldots+a_{D_0}}{D_0})^2] - (\mathbb{E}[a_n])^2$. For i.i.d traffic, this variance can be further evaluated and shown to equal $\frac{\sigma_a^2}{D_0}$. The total transmit power can now be approximated as,

$$\begin{aligned} P &= \frac{\sigma^2}{\sqrt{2\pi\sigma_a^2/D_0}} \int_{-\infty}^{\infty} (e^{Rx} - 1) e^{-\frac{(x-\lambda)^2 D_0}{2\sigma_a^2}} dx \\ &= \frac{\sigma^2}{\sqrt{2\pi\sigma_a^2/D_0}} \int_{-\infty}^{\infty} e^{-\frac{x^2+\lambda^2-2\lambda x-2R\sigma_u^2 x}{2\sigma_u^2}} dx - 1 \\ &= \sigma^2 \left(\frac{e^{R\lambda+R^2\sigma_u^2/2}}{\sqrt{2\pi\sigma_a^2/D_0}} \int_{-\infty}^{\infty} e^{-\frac{(x-\lambda-\sigma_u^2)^2}{2R\sigma_u^2}} dx - 1 \right) \\ &= \sigma^2 \left(e^{R\lambda} e^{\frac{R^2\sigma_a^2}{2D_0}} - 1 \right) \end{aligned}$$

Thus,

$$\log\left(1 + \frac{P}{\sigma^2}\right) = R\lambda + \frac{R^2\sigma_a^2}{2D_0}. \tag{16}$$

Note the remarkable similarity in the expression on the right and the form for the effective bandwidth [21]. The effective bandwidth is defined [21] as $\alpha(s,t) = \frac{1}{st}\log\mathbb{E}[e^{sX[0,t]}], 0 < s,t < \infty$. Indeed $\lambda R + \frac{R^2\sigma_a^2}{2D_0}$ exactly equals the effective bandwidth of a Gaussian source evaluated at $t = 1$, $s = \frac{1}{D_0}$. The Gaussian traffic source is defined [21] as $X[0,t] = \lambda R t + Z(t)$, where $Z(t)$ is normally distributed with zero mean. As expected, at $D_0 \to \infty$, the variation in input traffic are smoothened and we obtain Shannon's capacity formulation. Note also the similarity of (16) with the results in [8], which gives the average power as a function of the average delay. In [8], no bound on the absolute delay is imposed.

In the case of dependent traffic arrivals, we continue to assume that the output distribution is asymptotically Gaussian. In this case, the mean of the output distribution equals λ and the variance equals

$$\begin{aligned} \sigma_u^2 &= \mathbb{E}[u_n^2] - \mathbb{E}[u_n]^2 && (17) \\ &= \frac{1}{D_0^2}\mathbb{E}[(a_n + a_{n+1} + \ldots + a_{n+D_0-1})^2] - \lambda^2 && (18) \\ &= \sum_{i=0}^{D_0-1} \frac{(D_0 - i)}{D_0^2} R_a(i) - \lambda^2, && (19) \end{aligned}$$

where $R_a(i) = \mathbb{E}[a_n a_{n+i}]$ is the autocorrelation of the input sequence. Further, proceeding as before an approximate relationship for power can be derived as (11).

C. Closed form approximation: Statistical Scheduler

In the case of the statistical delay guarantees, we make the same assumption on Gaussianity of output traffic. In addition, we assume that the dropping policy is such that all high rate transmissions are truncated. Thus, the output traffic, which is assumed to have mean λ and variance σ_u^2, is truncated beyond size β. The truncation threshold β is calculated, as follows, to ensure that only δ fraction of packets are dropped

$$\begin{aligned} & \frac{1}{\sqrt{2\pi\sigma_u^2}} \int_{-\infty}^{\beta} e^{-\frac{(x-\lambda)^2}{2\sigma_u^2}} dx = 1 - \delta && (20) \\ \implies & \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\beta-\lambda}{\sigma_u}} e^{-s^2/2} ds = 1 - \delta && (21) \\ \implies & \Phi\left(\frac{\beta - \lambda}{\sigma_u}\right) = 1 - \delta, && (22) \end{aligned}$$

where Φ is the standard normal CDF. The transmission power can now be calculated as,

$$\begin{aligned} P &= \sigma^2 \frac{1}{\sqrt{2\pi\sigma_a^2/D_0}} \int_{-\infty}^{\beta} (e^{Rx} - 1) e^{-\frac{(x-\lambda)^2 D_0}{2\sigma_a^2}} dx \\ &= \sigma^2 \left(\frac{e^{R\lambda+R^2\sigma_u^2/2}}{\sqrt{2\pi\sigma_a^2/D_0}} \int_{-\infty}^{\beta} e^{-\frac{(x-\lambda-R\sigma_u^2)^2}{2\sigma_u^2}} dx - 1 \right) \\ &= \sigma^2 \left(e^{R\lambda} e^{\frac{R^2\sigma_a^2}{2D_0}} \Phi[\Phi^{-1}(1-\delta) - R\sigma_u] - 1 + \delta \right) \end{aligned}$$

By rearranging terms and taking logarithm, we obtain (12).

D. Optimality of statistical NOMP scheduler.

Heuristic Argument: For a delay bound of 1 time-slot, the proof follows directly from the exponential relationship between power and rate given by Shannon's Gaussian formula.[7]

[7]Actually a convex relationship is sufficient for the proof and all practical coding and modulation schemes follow a convex relationship.

Let each packet be subdivided into "bits" (or in general symbols) which are the smallest indivisible pieces of information. Let us assume that only one bit of information may be dropped from the entire sequence. Clearly, the optimum bit to drop that results in largest reduction of power is from the instant when u_n is maximum. Now, to drop the second (and all further) bits, a similar process can be used and thus bits are dropped from the largest resulting outputs. Essentially, this is a process of *inverse water-filling*, in which a threshold for the number of atoms (packets) is set and all packets bigger than the threshold are dropped. The dropping threshold is calculated to ensure that no more than $\delta\%$ of the packets are dropped.

A similar argument can be used for arbitrary delay bounds D_0. Initially compute the optimal NOMP scheduler. Now, if only one bit can be dropped, the optimal bit to drop is drop is from the largest occurring packet transmission in the deterministic NOMP scheduler. Repeating the process till $\delta\%$ of packets are dropped we see that the proposed stationary NOMP scheduler is optimal.

Proof: The problem of providing statistical guarantees is now posed as follows. Let u'_n denote the number of packets that would have been scheduled for transmission in time-slot n using the NOMP scheduler, *i.e.* with no packets violating the delay bound. Let u_n denote the number of packets transmitted by the statistical NOMP scheduler. Since, $\delta\%$ of the packets may violate the delay bound, we have the constraint $\sum_{n=1}^{N} u_n = (1-\delta)\sum_{n=1}^{N} u'_n \approx (1-\delta)\sum_{n=1}^{N} a_n$. The approximation $\sum_{n=1}^{N} u'_n \approx \sum_{n=1}^{N} a_n$, arises due to the boundary conditions[8] and does not affect the results for large N. The problem of minimizing the power is now posed as follows:

$$\min_{\{u_n\}} \quad \sigma^2 \sum_{n=1}^{N} e^{Ru_n} - 1 \tag{23}$$

$$\text{s.t.} \quad u_n \le u'_n \,, \sum_{n=1}^{N} u_n = (1-\delta)\sum_{n=1}^{N} u'_n \approx (1-\delta)\sum_{n=1}^{N} a_n$$

The solution to this optimization problem is easily derived using Lagrangian techniques and is given by $u_n = \min(u'_n, u_{dr})$, where u_{dr} is a dropping threshold that is calculated from $\sum_{n=1}^{N}(u'_n - u_{dr})^+ = (\delta)\sum_{n=1}^{N} a_n$.

REFERENCES

[1] S. Keshav, *An Engineering Approach to Computer Networks*. Addison-Wesley Longman, Inc., 1997.

[2] E. Biglieri, J. G. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2619–2692, October 1998.

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

[4] A. Ephremides and B. Hajek, "Information theory and communication networks: An unconsummated union," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2416–2434, Oct. 1998.

[5] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inform. Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[6] D. J. Goodman, J. Borras, N. B. Mandayam, and R. Yates, "INFOSTATIONS: A new system model for data and messaging services," *Proc. Vehicular Technology Conference*, pp. 969–973, May 1997.

[7] B. Collins and R. L. Cruz, "Transmission policies for time varying channels with average delay constraints," in *Proc. Allerton Intnl. Conf. on Comm., Control and Computing*, Monticello, IL, 1999, pp. 709–717.

[8] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay bounded packet scheduling of bursty traffic over wireless channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 125–144, 2004.

[9] B. Prabhakar, E. U. Biyikoglu, and A. E. Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Proc. INFOCOM*, Anchorage, Alaska, April 2001.

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

[11] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proceedings of the IEEE*, October 1995.

[12] V. Bharghavan, S. Lu, and T. Nandagopal, "Fair scheduling in wireless packet networks: Issues and approaches," *IEEE Pers. Commun. Mag.*, vol. 6, no. 1, pp. 44–55, Feb. 1999.

[13] X. Liu, E. K. P. Chong, and N. B. Shroff, "Transmission scheduling for efficient wireless utilization," *Proc. INFOCOM*, 2001.

[14] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "CDMA data QoS scheduling on the forward link with variable channel conditions," *Tech. Rep., Bell Labs.*, 2000.

[15] N. Joshi, *et. al*, "Downlink scheduling in CDMA data networks," in *Proc. MOBICOM*, 2000, pp. 179–190.

[16] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.

[17] R. Knopp and P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. ICC*, 1995.

[18] D. Tse, "Opportunistic communications: Smart scheduling and dumb antennas," *Seminar presented at Intel Corp.*, January 2002.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, New York, 1991.

[20] D. Rajan, A. Sabharwal, and B. Aazhang, "Outage behaviour with delay and CSIT," in *Proc. ICC*, Paris, France, June 2004, pp. 578–582.

[21] F. Kelly, "Notes on effective bandwidth," *Technical report available at http://www.statslab.cam.ac.uk/frank/*.

[22] "Thirty days of wide-area tcp connections," *http://ita.ee.lbl.gov/html/contrib/LBL-CONN-7.html*.

[23] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE Trans. Networking*, vol. 3, no. 3, pp. 226–244, June 1995.

[24] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic," *Proc. ACM SIGCOMM*, 1993.

[25] S. Lin and D. J. Costello, *Error control coding: Fundamentals and Applications*. Prentice Hall, second edition, 2004.

[8] all a_N packets may not be transmitted during time-slot N.

Dinesh Rajan received the B.Tech. degree in Electrical Engineering from Indian Institute of Technology, Madras in 1997. He received his M.S. and Ph.D. degrees in Electrical and Computer Engineering in 1999 and 2002, respectively, from Rice University, Houston, Texas. He joined the Electrical Engineering Department at Southern Methodist University, Dallas, Texas in 2002, where he is currently an assistant professor. He received a NSF CAREER award in 2006. His current research interests include communications theory, wireless networks and information theory.