On Scheduling Transmissions under QoS based Constraints

Premal Shah *and* ***Abhay Karandikar***

Information Networks Lab,
Department of Electrical Engineering,
IIT Bombay
Mumbai, India

June 2003

Outline

- The QoS Framework and the Token Bucket Regulator
- The Regulated Media Streaming Problem
- Information Utility of a Token Bucket Regulator
- Power-efficient Transmissions and Discrete Rate Scheduling

QoS Framework and the Token Bucket Regulator

- **Service Level Agreement (SLA) between the Network and the Source**
  - Delay and loss performance guarantees for flows
  - Bounds on Source Traffic Profile
- **Linearly Bounded Arrival Processes**
  - Maximum amount of traffic $A(t, \tau)$ that a source may send in an interval $(t, \tau]$ is bounded by a linear function of the interval
  - $(\rho - \sigma)$ regulated traffic

$$A(t, \tau) \quad \leq \quad \sigma + \rho(\tau - t)$$

Token Bucket Regulator

- **The standard IETF Token Bucket Regulator**
  2 parameters
  - Token Refill Rate - r
  - Bucket Depth (maximum no. of residual tokens)- B
- **Augmented Token Bucket Regulator** $TBR(\boldsymbol{r}, \boldsymbol{B}, B_0)$
  - Token Refill Rate vector - $\boldsymbol{r}$: $(r_1, r_2, ..., r_N)$
  - Bucket Depth vector - $\boldsymbol{B}$: $(B_1, B_2, ..., B_N)$
  - Initial Token Grant - B_0

Token Bucket Regulation Constraints

- packet length vector - $\boldsymbol{x}$: $(x_1, x_2, ..., x_N)$
- residual token vector - $\boldsymbol{t}$: $(t_1, t_2, ..., t_N)$

for a conforming flow

$$\begin{aligned} x_n &\leq t_{n-1} + r_n \ \ ;\forall n : 1 \leq n \leq N \\ t_n &= \min(t_{n-1} + r_n - x_n, B_n) \ \ ; \\ t_0 &= B_0 \end{aligned} \tag{1}$$

Admissibility Constraints (Necessary and Sufficient)

$$\sum_{n=i}^{n=j} x_n \leq B_{i-1} + \sum_{n=i}^{n=j} r_n \quad \forall (i,j) : 1 \leq i \leq j \leq N \tag{2}$$

Distortion Model for Streaming

Stream is assumed to be characterized by a requirement schedule $\boldsymbol{y} : (y_1, y_2, y_3, ..., y_N)$

- y_n : bits required to code media content for interval n
- u_n : no. of bits in the transmitted stream for coding interval n
- $\alpha_n = \max(y_n - u_n, 0)$: absolute loss for interval n
- $\beta_n = \max(\frac{y_n - u_n}{y_n}, 0)$: fractional loss for interval n
- d_n : distortion for interval n; a **convex, increasing** non-time varying function of α_n or β_n

Modeling Overall Distortion

Assume that the overall distortion may be modeled by

$$D = \sum_{n=1}^{n=N} d_\alpha(\alpha_n) \text{ OR} \quad (3)$$

$$= \sum_{n=1}^{n=N} d_\beta(\beta_n) \text{ OR} \quad (4)$$

$$= \sum_{n=1}^{n=N} y_n . d_\beta(\beta_n) \quad (5)$$

Justification for Distortion Model

- Content in each interval made up of objects
- Each object follows a Gaussian distribution
- An object needs to be coded using M bits

for optimal performance

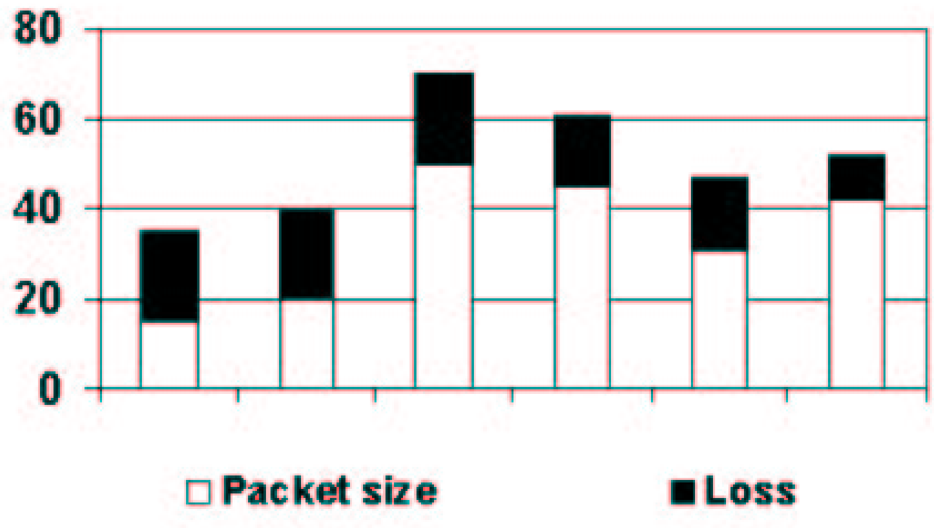
- Each object coded using $M\frac{u_n}{y_n}$ bits
- Distortion (mean squared error sense) in the coding of each object:
  $e = (2^{2M\beta_n} - 1)$
- Distortion d_n for the interval n:
  $d_n = \frac{y_n}{M}(2^{2M\beta_n} - 1)$

thus, *Overall distortion* : $D = \frac{1}{M}\sum_{n=1}^{n=N} y_n(2^{2M\beta_n} - 1)$

An Optimization Cost Function

- Assumptions on the Distortion Function:
  - Convexity
  - Additivity
- Objective: *Minimize overall distortion and distortion jitters*
- Optimization Cost function: β' *- vector of fractional losses arranged in a decreasing order*
- Offline solutions optimal w.r.t. β'
  - have minimum MINMAX distortion
  - are optimal (for 5)) and close to optimal (for (4)) in terms of overall distortion

Offline Packet Scheduling



To optimize β':

- Distribute losses β_n equitably
- Begin with assigning β_N

Properties of Optimal Schedule

If no delay is allowed in transmissions then $u_n = x_n$.
Define

$$\begin{aligned}
\gamma_{ij} &= \max\{\frac{\sum_{n=i}^{n=j}[y_n - r_n] - B_{i-1}}{\sum_{n=i}^{n=j} y_n}, 0\} \\
\gamma^* &= \max_{(i,j):1\leq i\leq j\leq N}\{\gamma_{ij}\} \\
(i^*, j^*) &= \text{arg} \max_{(i,j):1\leq i\leq j\leq N}\{\gamma_{ij}\}
\end{aligned}$$

In an allocation that is optimal w.r.t β'

$$\begin{aligned}
\beta'_1 &= \gamma^* \\
\beta_n &= \gamma^* \;\; ;\forall n: \;\; i^* \leq n \leq j^* \\
x_n &= y_n(1-\gamma^*)
\end{aligned}$$

Non-zero Absolute Delay Bounds

- x_n now denotes the number of bits allocated to code content of interval n
- A delay of D intervals is allowed before transmission
- Zero delay analysis carries over with few modifications:

$$\sum_{n=i}^{n=j} x_n \leq B_{i-1} + \sum_{n=i}^{\min(j+D,N)} r_n \quad \forall (i,j) : 1 \leq i \leq j$$

$$\gamma_{ij} = \max\{\frac{\sum_{n=i}^{n=j} y_n - \sum_{n=i}^{\min(j+D,N)} r_n] - B_{i-1}}{\sum_{n=i}^{n=j} y_n}, 0\}$$

Online Packet Scheduling

- Optimal solution for expected distortion - *Dynamic Programming*:
  - requires knowledge of the distortion function
  - high complexity
- Develop simpler schemes
- Use offline Programming loss property:

$$\beta_i \leq \max_{j:i \leq j \leq N} \{\{1 - \frac{t_{i-1} + (j-i+1)r}{\sum_{n=i}^{n=j} y_n}\} \cup \{0\}\}$$

- Equality holds for sufficiently large B

A Heuristic Online Packet Scheduling Algorithm

- Replace β_i by $E[\beta_i]$ conditioned on known inputs
- Assumptions:
  - input stream is Stationary
  - 1^{st} order Markov dependence between y_is
  - B is large
  - certain other approximations
- Online policy:

$$\beta_i(t, y) = \max_{q: 0 \le q \le N-i} \{1 - [t + (q+1)r] E[R_{0q}(y)]\}$$

$$R_{ij}(y_i) = \frac{1}{y_i + \sum_{n=i+1}^{n=j} y_n} \quad \forall i \le j \le N$$

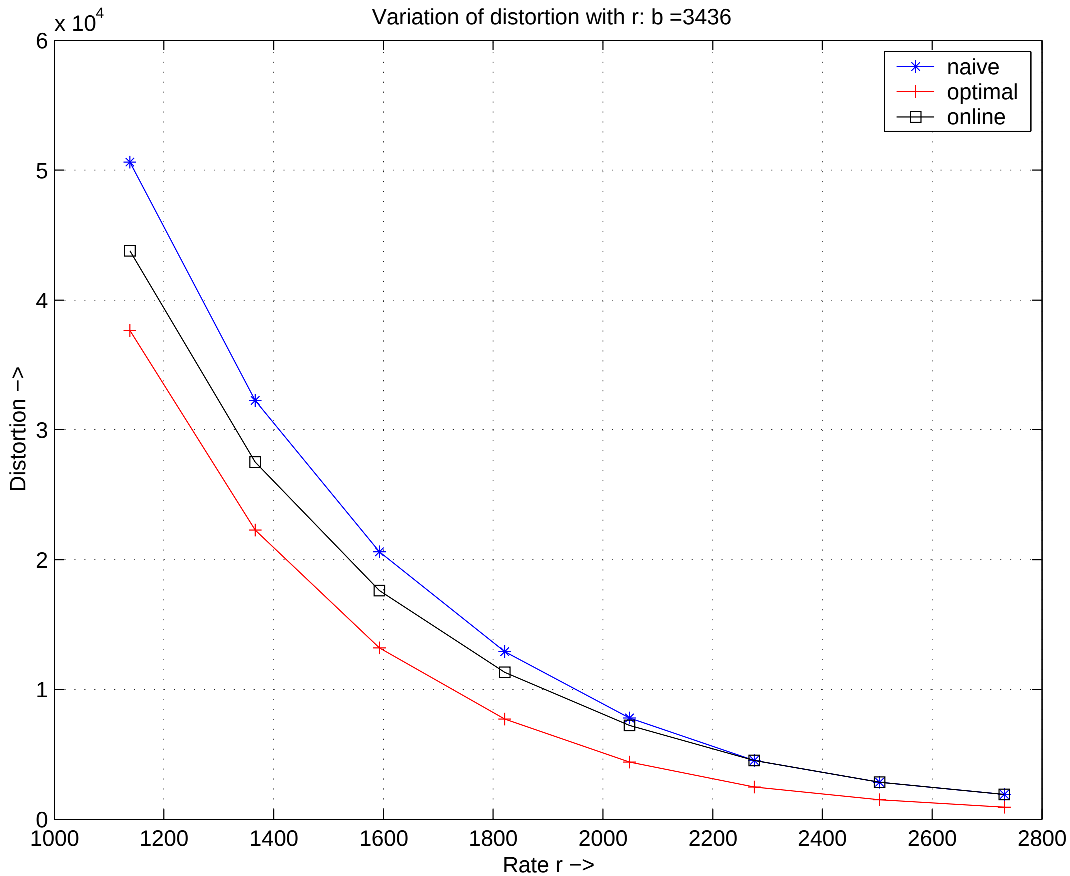
Simulation Results

Comparison of Distortion performance of Various Algorithms (Weighted Fractional Measure)

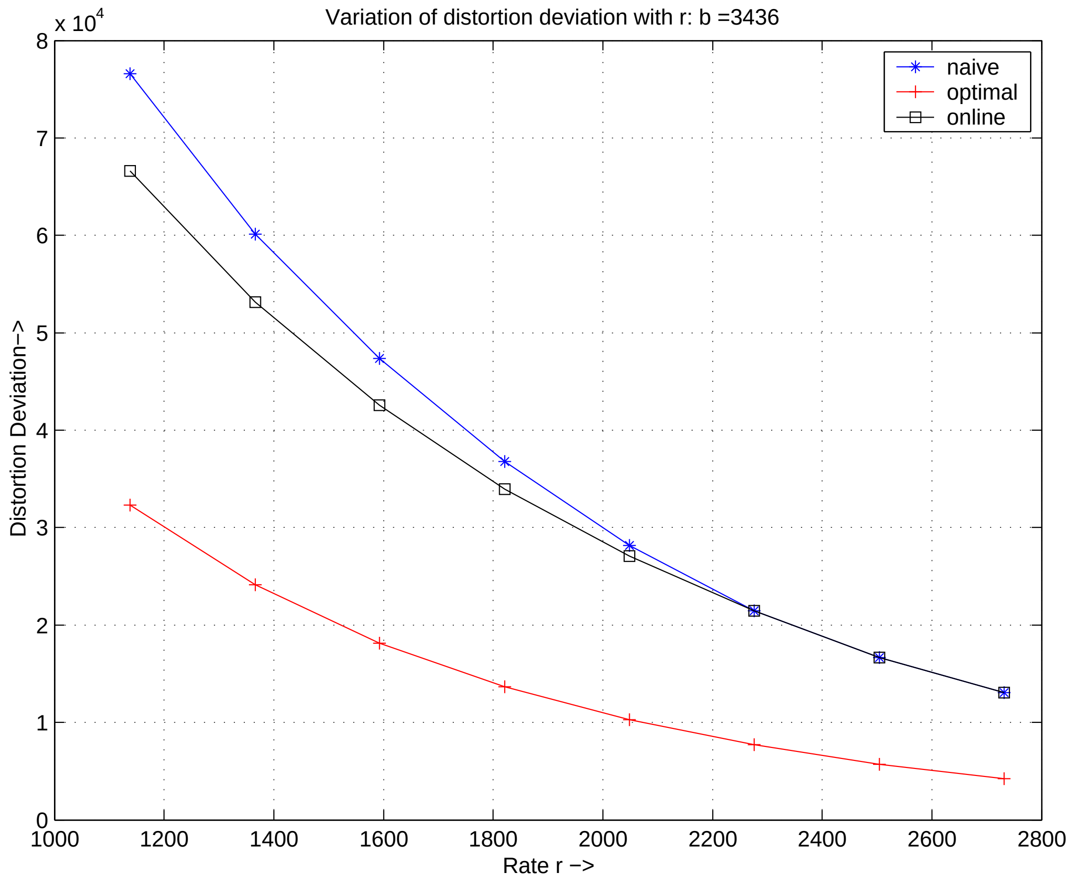
- $r \approx 0.8 * E[y_n]$
- $B \approx 4 * S.D(y_n)$

Trace	Distortion			S.D. in Distortion		
	Opt.	Naive	Online	Opt.	Naive	Online
bean(MPEG4)	1	1.3030	1.2535	1	2.0349	1.9547
bean(var)	1	1.6788	1.4679	1	2.6964	2.4908
formula(MPEG4)	1	1.2646	1.1441	1	1.5179	1.2939
formula(var)	1	1.5377	1.3353	1	2.0724	1.8620

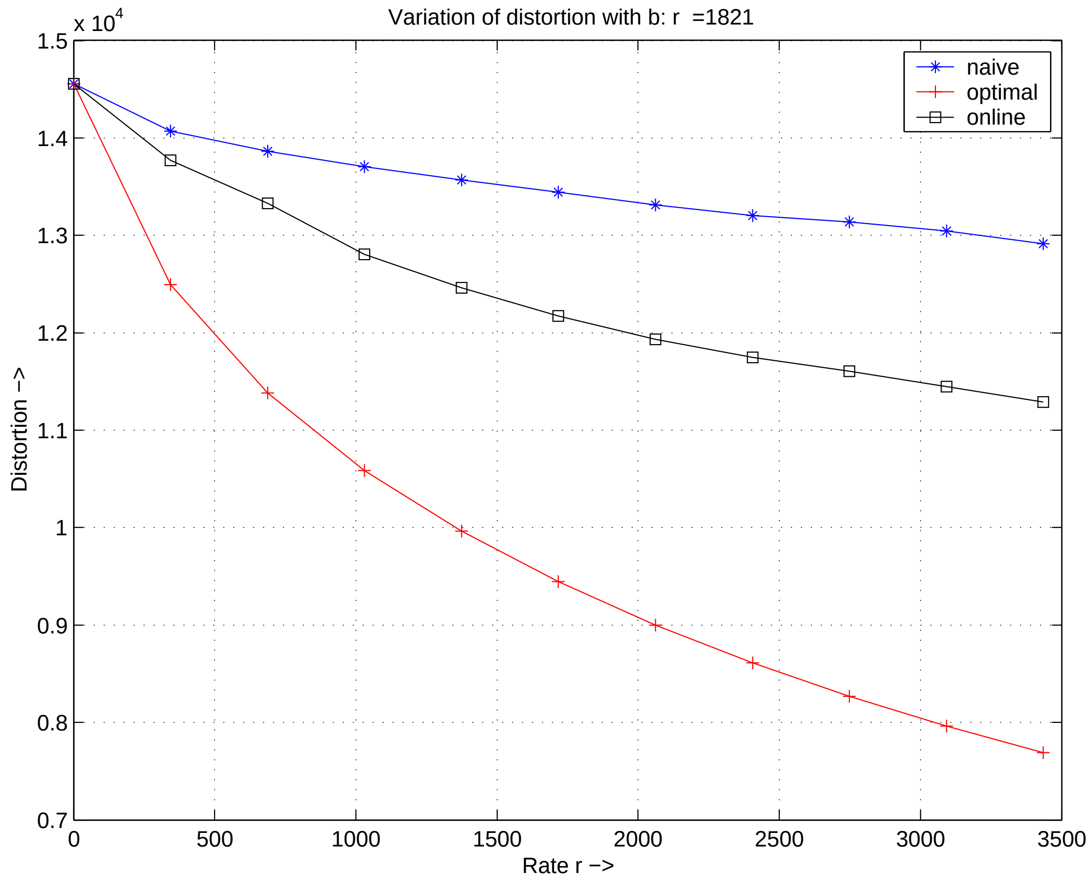
Dependence of Distortion on r



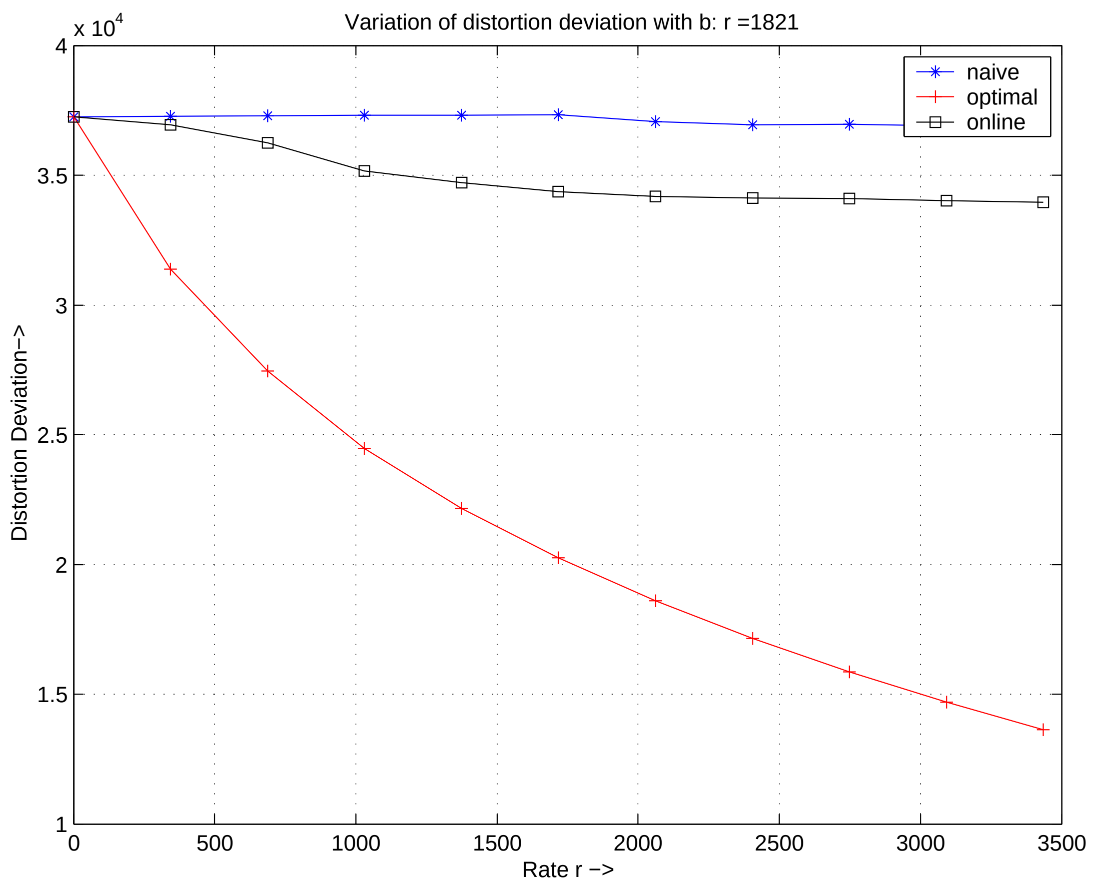
Dependence of Distortion Deviation on r



Dependence of Distortion on B



Dependence of Distortion Deviation on B



Conclusions and Future Directions

- *Packet length scheduling* - a useful technique to obtain better distortion performance for constrained streaming
- Performance improvements more marked for the offline case and for *small values of* r and *large values of* B
- More efficient online scheduling techniques
- Joint selection of token bucket parameters and scheduling to optimize a joint criterion based on price and distortion

Side Information Channels

- Information can be conveyed through means other than packets themselves, such as timing of packets
- Information in ‘side information’ channels may be distorted by a network randomizing delays in packet transmission
- In a QoS Network that offers loss and delay guarantees, the ‘side information’ channels become distortion-free for conforming flows

Token Bucket Regulator Constrained Flow

The flow has two ways of conveying information -

- Packet contents
- Packet lengths that can be varied within the limits allowed by the regulator

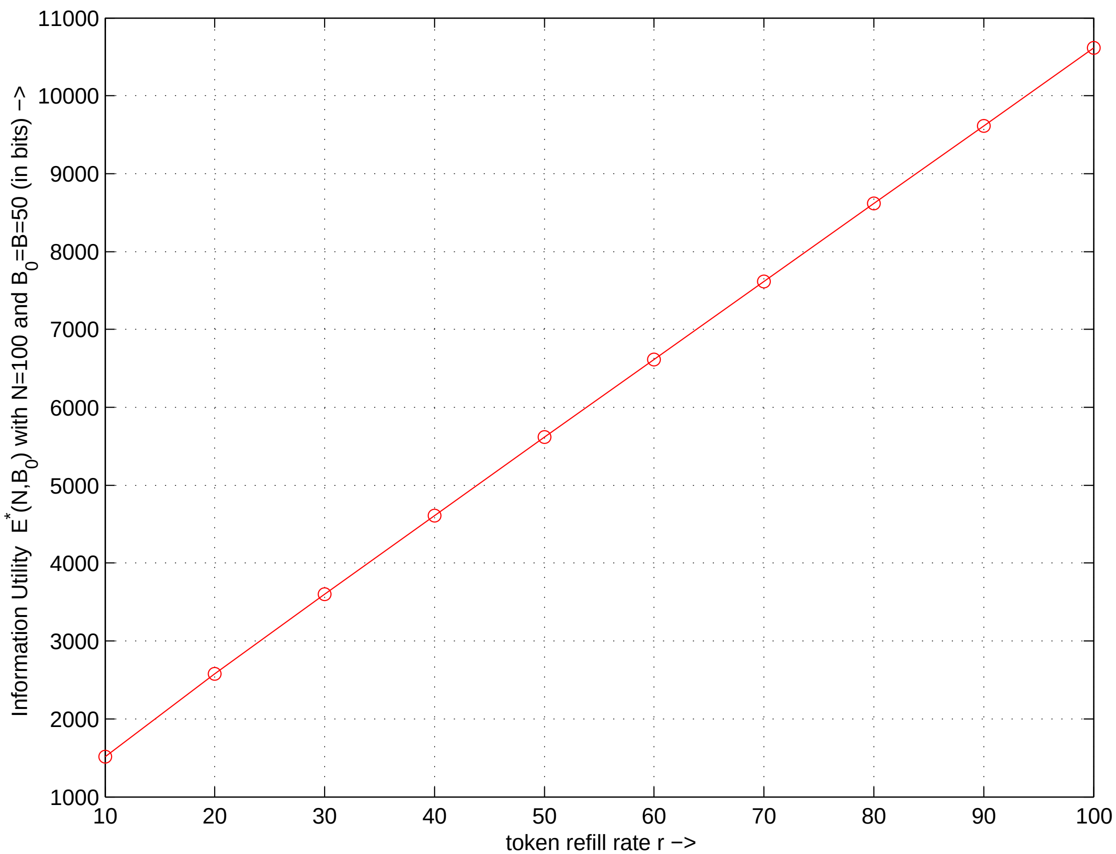
Maximum average information of a duration N flow

- b: no. of residual tokens
- n: No. of transmissions to be made
- $\mathcal{E}^*(b,n)$: Maximum entropy of a duration n flow
- $p_i^*(b,n)$: Optimal probability of choosing a packet of length i

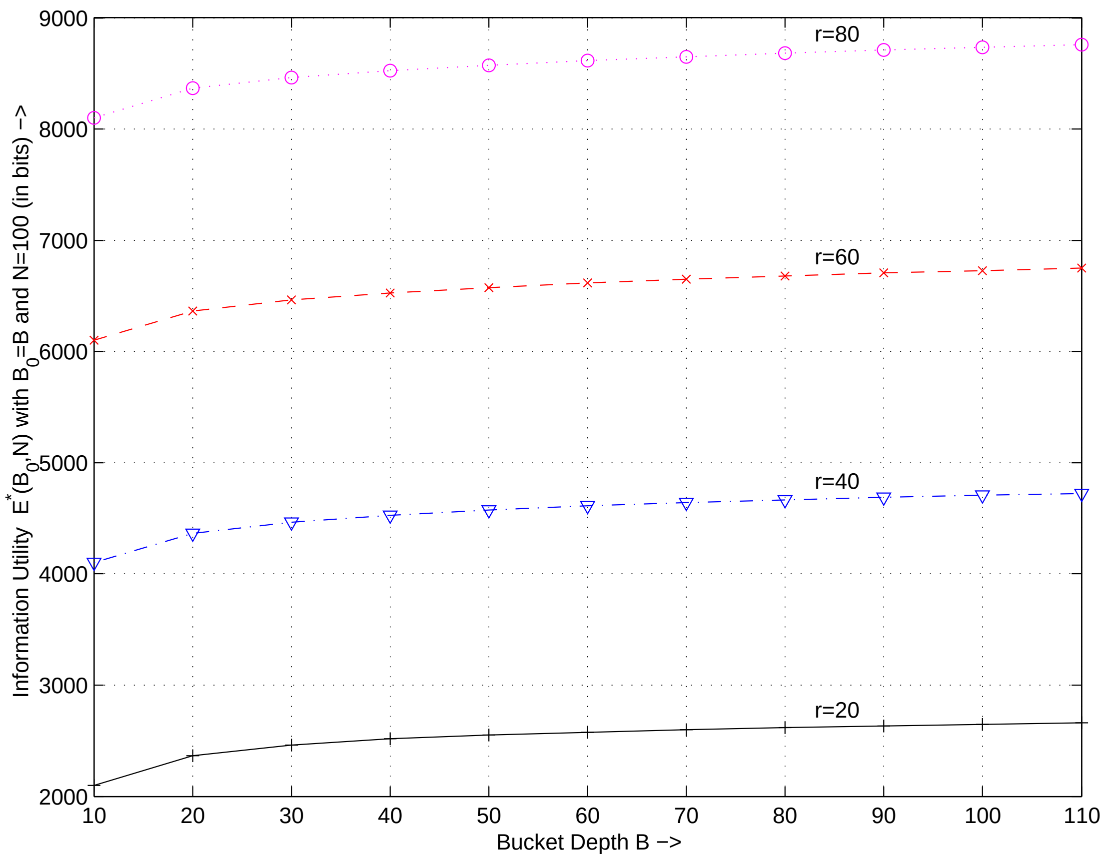
$$\mathcal{E}(b,n) = \sum_{i=0}^{i=b+r} p_i(b,n)[i \ln 2 - \ln p_i(b,n) + \mathcal{E}(b+r-i, n-1)]$$

$$p_i^*(b,n) \propto e^{i \ln 2 + \mathcal{E}^*(b+r-i, n-1)}$$

Dependence of Information Utility on r



Dependence of Information Utility on B



Comments and Future Directions

- Pricing will generally be linear in both r and B
- Information Utility of a flow under Token Bucket Regulation increases *linearly with* r but *sub-linearly with* B
- Evaluating the Information Utility under specific loss and delay guarantees

Scheduling Transmissions for Power Efficiency

- Transmitter - a principal consumer of power, a key performance metric for wireless systems
- Two means of delay-power tradeoff possible in wireless channels
- Time varying nature of the channel
- Convex nature of the Power-Rate ($P - R$) relation

Power saving: Time Varying Nature of the Channel

- The wireless channel experiences periods of fades
- Transmissions at a particular rate during fades require much higher power than what would be required on an average
- Defer transmissions of packets arriving during fades to periods when the channel state is good
- Average Power consumption decreases at the cost of increased average delay

Power saving: Convex $P - R$ relation

- The Shannon capacity relation for an AWGN channel

$$P = \sigma^2(e^R - 1) \tag{6}$$

- Practical communication schemes also show a convex $P - R$ relation
- Gains in power savings possible by transmitting in appropriately sized batches.
- Buffering delay increases as a result
- Holds even for wired networks

Example: Batch size scheduling for Power saving

- Suppose power is given by (6) with $\sigma = 1$. Consider an arrival process with r arrivals every alternate slot and the two transmission schemes -

  Scheme A: Transmit as arrive

$$P_A = \frac{e^r - 1}{2}$$
$$D_A = 0$$

  Scheme B: Transmit $\frac{r}{2}$ packets in every slot

$$P_B = e^{\frac{r}{2}} - 1$$
$$D_B = \frac{1}{2}$$

- Delay budget and the arrival statistics limit the gains possible

Discrete Rate Scheduling

- **Objective:** Minimize average power consumption subject to an average delay constraint $\bar{d}$
- Why Discrete Rate Schdeuling?
  Transmission at arbitrary rates demands
  - high transmitter complexity
  - a significant signaling overhead and receiver complexity

Discrete Rate Schedulers or Constant Power Schedulers $(\mathcal{P})$-
Transmit only at from a finite set of pre-assigned rates

System Model

- Discrete time (slotted) with constant channel conditions.
- iid packet arrival process a_n

$$\begin{aligned} P[a_n = i] &= p_i \\ E[a_n] &= \lambda \end{aligned}$$

- Stationary scheduling policy (u_n)
  Buffer length evolves as

$$x_{n+1} = x_n - u_n + a_n$$

Greedy Schedulers

K**-rate Scheduler** $(\mathcal{P}^K)$**:** Transmit from a set $\mathcal{N}^K : \{N_1, N_2, ..., N_K\}$ of K- non-zero rates

Monotone Scheduler: A deterministic Scheduler, where scheduling function $u(.)$ is increasing in buffer occupancy x

Greedy Scheduler: A deterministic Scheduler that transmits as many as it can

The Greedy Scheduler achieves the best delay performance of the schedulers allowed to transmit at from the same set of rates

Optimal 1-rate scheduler

For 1-rate schedulers

- Average power consumption is independent of scheduling policy and depends only on rate N

$$C_N = \frac{\lambda}{N} E(N) \tag{7}$$

- C_N increases with N

Optimal scheduler is a greedy scheduler transmitting at rate N^* (G_{N^*}), where N^* is the smallest N for which average delay $< \bar{D}$

Queueing Analysis of a Greedy 1-rate Scheduler

Steady state flow balance equations for G_N,

$$\pi_i = \sum_{j=0}^{j=N-1} \pi_j p_{i-j} + \sum_{j=N}^{j=\infty} \pi_j p_{i-j+N}$$

Using transfer functions these can be written as,

$$\begin{aligned} \boldsymbol{\pi}(z) &= \boldsymbol{p}(z)\boldsymbol{\pi}^{\mathbf{0}}(z) + \boldsymbol{p}(z)\frac{\boldsymbol{\pi}(z) - \boldsymbol{\pi}^{\mathbf{0}}(z)}{z^N} \quad \text{or,} \\ \boldsymbol{\chi}(z) &= \frac{\boldsymbol{p}(z) - 1}{z^N - \boldsymbol{p}(z)}\boldsymbol{\pi}^{\mathbf{0}}(z) \end{aligned} \tag{8}$$

where, $\boldsymbol{\pi}^{\mathbf{0}}(z) = \sum_{i=0}^{i=N-1} \pi_i z^i$ and $\boldsymbol{\chi}(z) = \frac{\boldsymbol{\pi}(z) - \boldsymbol{\pi}^{\mathbf{0}}(z)}{z^N}$

- When $N > \lambda$ the queue is stable and $\boldsymbol{\chi}(z)$ must converge
- $\boldsymbol{\pi}^{\mathbf{0}}(z)$ is a polynomial of degree $N-1$

Greedy 1-rate Scheduling: Geometric Arrival Process

$$\boldsymbol{p}(z) = \frac{1-\alpha}{1-\alpha z}$$

$$\boldsymbol{\chi}(z) = \frac{\alpha(z-1)}{(z^N-1)-\alpha(z^{N+1}-1)}\boldsymbol{\pi}^{\mathbf{0}}(z)$$

When $N > \lambda$, $(z^N - 1) - \alpha(z^{N+1} - 1)$ has a root $\beta > 1$

$$\boldsymbol{\chi}(z) = \frac{1-\beta^{-1}}{N(1-\alpha)}\frac{1}{1-\beta^{-1}z}$$

$$\boldsymbol{\pi}(z) = \frac{1-\beta^{-1}}{N}\frac{\sum_{i=0}^{N-1} z^i}{1-\beta^{-1}z}$$

$$D_N = \frac{1-\alpha}{\alpha}[\frac{N-1}{2}+\frac{1}{\beta-1}]$$

Greedy 1-rate Scheduling: Polynomial Arrival Processes

- $\boldsymbol{p}(z)$ - a polynomial of degree $R \geq N$, and $p_0 > 0$
- $z^N - p(z) = c(z-1)\boldsymbol{g}(z)\boldsymbol{h}(z)$
- $\boldsymbol{h}(z)$ has zeros outside the unit circle only and $\boldsymbol{g}(z)$, only on or inside the unit circle

If ($N > \lambda$), Rouche's theorem $\Rightarrow$

- $\boldsymbol{g}(z)$ is of degree $N-1$
- $\boldsymbol{g}(z)$ has zeros within the unit circle only

Greedy 1-rate Scheduling: Polynomial Arrival Processes (cont'd)

Then,

$$\begin{aligned}
\boldsymbol{\pi}^{\mathbf{0}}(z) &= \frac{c\boldsymbol{h}(1)}{N}\boldsymbol{g}(z) \\
\boldsymbol{\pi}(z) &= \frac{\boldsymbol{h}(1)}{N}\frac{\boldsymbol{p}(z)\sum_{i=0}^{i=N-1} z^i}{\boldsymbol{h}(z)} \\
D_N &= 1 + \frac{N-1}{2\lambda} - \frac{\boldsymbol{h}'(1)}{\boldsymbol{h}(1)\lambda}
\end{aligned}$$

Greedy Multirate Scheduling

K-rate scheduler with $\mathcal{N}^K = \{N_1, N_2, ..., N_K\}$

Steady state equation

$$\sum_{i=0}^{i=K-1} z^{N_i} \boldsymbol{\pi^i}(z) + z^{N_K} \boldsymbol{\chi}(z) = \boldsymbol{p}(z)[\sum_{i=0}^{i=K-1} \boldsymbol{\pi^i}(z) + \boldsymbol{\chi}(z)] \quad (9)$$

$$\text{where, } \boldsymbol{\pi^i}(z) = \sum_{j=N_i}^{j=N_{i+1}-1} \pi_j z^{j-N_i}$$

$$\boldsymbol{\chi}(z) = \sum_{j=N_K}^{j=\infty} \pi_j z^{j-N_K}$$

Greedy Multirate Scheduling: Geometric Arrival Process

From (9),

$$\boldsymbol{\chi}(z) \quad = \quad \frac{\alpha(z-1)\boldsymbol{\pi}^{\mathbf{0}}(z) + \sum_{i=0}^{i=K-1}[\alpha(z^{N_i+1}-1) - (z^{N_i}-1)]\boldsymbol{\pi}^{\boldsymbol{i}}(z)}{(z^{N_K}-1) - \alpha(z^{N_K+1}-1)}$$

If $(N_K > \lambda)$, then $(z^{N_K}-1) - \alpha(z^{N_K+1}-1) = -\alpha(z-1)(z-\beta)g(z)$, where $\beta > 1$, and $\boldsymbol{g}(z)$ has zeros inside the unit circle $\Rightarrow$

$$\boldsymbol{\chi}(z) \quad = \quad -\frac{A}{z-\beta}$$

$$\boldsymbol{\pi}^{\mathbf{0}}(z) + \sum_{j=1}^{j=K-1} \boldsymbol{\pi}^{\boldsymbol{j}}(z)[\sum_{i=0}^{i=N_j} z^i - \frac{1}{\alpha}\sum_{i=0}^{i=N_j-1} z^i] \quad = \quad A\boldsymbol{g}(z)$$

- $\boldsymbol{\pi}^{\boldsymbol{i}}(z)$ - obtained from repeated Euclidean divisions
- A can be obtained using $\boldsymbol{\pi}(1) = 1$

Greedy Multirate Scheduling: Polynomial Arrival Process

From (9),

$$\boldsymbol{\pi}(z) = \boldsymbol{p}(z)\frac{z^{N_K}\widetilde{\boldsymbol{\pi_l}}(z) - \boldsymbol{\pi_l}(z)}{z^{N_K} - \boldsymbol{p}(z)}$$

$$\text{where, } \widetilde{\boldsymbol{\pi_l}}(z) = \sum_{i=0}^{i=K-1} \boldsymbol{\pi^i}(z)$$

If $(z^{N_K} - \boldsymbol{p}(z)) = c(z-1)\boldsymbol{g}(z)\boldsymbol{h}(z)$ and $N_K > \lambda$, then

$$\boldsymbol{\pi}(z) = \boldsymbol{p}(z)\frac{\boldsymbol{u}(z)}{c\boldsymbol{h}(z)}$$

- $\boldsymbol{u}(z)$ - unknown polynomial of the degree of $\widetilde{\boldsymbol{\pi_l}}(z))$
- obtained by solving a system of $(\text{deg}(\widetilde{\boldsymbol{\pi_l}}(z)) + 1)$ linear equations

K-rate Deterministic Monotone Scheduler

- Characterized by transmission rates $\mathcal{N}^K$: $\{N_1, N_2, ..., N_K\}$ and transmission thresholds $\mathcal{T}^K$: $\{t_1, t_2, ..., t_K\}$
- *Scheduling Action:* $u(x) \geq N_i \iff x \geq t_i$

Flow Balance Equation:

$$\boldsymbol{\pi}(z) = \boldsymbol{p}(z)\Big[\sum_{i=0}^{i=K-1} z^{t_i - N_i}\boldsymbol{\pi}^{\boldsymbol{i}}(z) + z^{t_K - N_K}\boldsymbol{\chi}(z)\Big]$$

$$\Rightarrow \boldsymbol{\chi}(z) = \frac{\boldsymbol{p}(z)\sum_{i=0}^{i=K-1} z^{t_i - N_i}\boldsymbol{\pi}^{\boldsymbol{i}}(z) - \sum_{i=0}^{i=K-1} z^{t_i}\boldsymbol{\pi}^{\boldsymbol{i}}(z)}{z^{t_K - N_K}\left(z^{N_k} - \boldsymbol{p}(z)\right)}$$

$$\text{where, } \widetilde{\boldsymbol{\pi}_{\boldsymbol{l}}}(z) = \sum_{i=0}^{i=K-1} z^{t_i - N_i}\boldsymbol{\pi}^{\boldsymbol{i}}(z)$$

Can be solved using techniques similar to the Greedy Scheduler case, with some modifications

Minimum Power Requirement

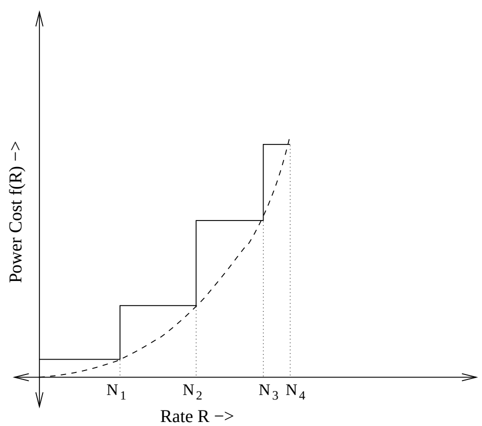
- Greedy Schedulers give the best delay performance
- The K-rate scheduler with the least power requirement (in the absence of delay constraints, i.e., $\bar{D} = \infty$) is characterized as follows

$$
\begin{aligned}
q_p^* &= \frac{\lambda - N_p}{N_{p+1} - N_p} \\
q_{p+1}^* &= \frac{N_{p+1} - \lambda}{N_{p+1} - N_p} \\
q_i^* &= 0 \,; i \neq p, p+1
\end{aligned}
$$

  - q_i^* : fractional time for which the optimal scheduler transmits at rate N_i,
  - p : integer such that $N_p < \lambda < N_{p+1}$
- $P_{min} = \frac{N_{p+1} - \lambda}{N_{p+1} - N_p} E(N_p) + \frac{\lambda - N_p}{N_{p+1} - N_p} E(N_{p+1})$

Future Directions

- Nature of the Optimal Deterministic Scheduler:
  Is it monotone?
- The pseudo-all rate transmitter:
  - Uses stuffing or adds dummy packets to flush packets
  - Breaks the Greedy Scheduler Delay bound
  - Modified cost function



- Scheduling for Time-varying Channels

Thank You