Fundamental Limits of Traffic Regulation and Their Connections to Source Coding

Cheng-Shang Chang[1] Chi-Chao Chao[2] Joy A. Thomas[3]

Abstract

In this paper, we explore the analogy between queueing theory and information theory, especially in traffic regulation and source coding. We consider a traffic regulation problem, where an input sequence is mapped to an output sequence. The rate of a regulator is defined to be the "peak" rate of the output sequence and the performance of a regulator is in terms of loss probability. Under our formulation, we show theorems that are analogous to Shannon's source coding theorem and the universal block coding theorem. For real-time traffic regulators that have a maximum delay constraint, the trade-off between delay and loss probability is characterized by the recently developed notion of effective bandwidth.

Key Words: traffic regulation, effective bandwidth, source coding.

1 Introduction

Queueing theory and information theory are two of the most important theories in Telecommunications. In view of the telecommunications hierarchy, information theory covers topics in the physical link level, such as source coding, channel coding, and so forth. The fundamental limits are usually in terms of the *entropy* rate (of an input) and the *average* rate (of an output). *Bit Error Rate* (BER) is the commonly used measure for the performance in a physical link. Unlike information theory, queueing theory deals with problems in the cell level or higher. Interesting performance measures include *Cell Loss Probability* (CLP) and *cell delay*. The fundamental limits of queueing theory, usually referred to the stability of queueing systems, are the *average* rate (of an input), and the recently developed notion of *effective bandwidth* (see e.g., [22, 24, 20, 4, 25, 5, 18, 33, 21, 26, 12, 11, 14, 23, 6], among many others.) For the theory of effective bandwidth, we refer to [33] for historical remarks and [6] for a review.

One of our main objectives in this paper is to explore the analogy between queueing theory and information theory. We start from defining a traffic regulation problem in Section 2. As in source coding, a traffic regulator takes an input sequence to an output sequence. The rate of a regulator is defined to be the "peak" rate of the output sequence. Analogous to Shannon's source coding theorem, we show that under our formulation the loss probability can be made arbitrarily small as long as the average rate of the input sequence is smaller than the rate of the regulator.

In Section 3, we further explore the connections between source coding and our traffic regulation problem. Analogous to block coding, we consider block regulators that take an input with block size n to an output with the same block size. We show the block loss probability decays exponentially at a rate characterized by the relative entropy (or the Kullback-Leibler distance) when the block size n is increased to infinity. The argument for the result is completely parallel to that in universal block coding.

The trade-off between delay and loss probability for traffic regulation is addressed in Section 4. We consider real-time regulators that have a maximum delay constraint. Cells that cannot meet the delay constraint are lost. We identify the exponential decay rate of loss probability as the delay constraint is relaxed to infinity. The decay rate is related to the recently developed notion of effective bandwidth.

We conclude the paper in Section 5 by discussing possible extensions such as relaxation of the rate constraint, relaxation of the input sequence, and the queueing analogy for channel coding. The mathematical tool that we use in this paper is the *large deviation principle* (LDP), such as Sanov's theorem and the Gärtner-Ellis theorem.

2 A traffic regulation problem

We start from defining the entropy rate, the average rate and the "peak" rate of stochastic processes.

Definition 2.1 (i) Let $\mathbf{X} = \{X_n, n \geq 1\}$ be a sequence of discrete random variables on the finite alphabet $\Sigma = \{a_1, \ldots, a_d\}$. The *entropy* rate of a sequence of random variables $\mathbf{X}$, denoted by $H_{etr}(\mathbf{X})$, is defined as follows:

$$H_{etr}(\mathbf{X}) = -\limsup_{n\to\infty} \frac{1}{n} \sum_{\omega\in\Sigma^n} \mathrm{P}_{\mathbf{X}}(\omega) \log \mathrm{P}_{\mathbf{X}}(\omega),$$

where $\mathrm{P}_{\mathbf{X}}(\omega) = \mathrm{P}((X_1, \ldots, X_n) = \omega)$ for $\omega \in \Sigma^n$.

[1]Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30043, Taiwan, R.O.C.; cschang@ee.nthu.edu.tw. Supported in part by the National Science Council, Taiwan, R.O.C., under Contracts NSC 83-0208-M007-091 and NSC 83-0404-E007-052.

[2]Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30043, Taiwan, R.O.C.; ccc@ee.nthu.edu.tw.

[3]IBM Research Division, T.J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598, U.S.A.; jat@watson.ibm.com

In particular, if $\mathbf{X}$ is a sequence of i.i.d. random variables with the probability density function $\mathrm{P}(X_1 = x_i) = q_i, \;\; i = 1, \ldots, d$, then

$$\begin{aligned} H_{etr}(\mathbf{X}) &= -E \log \mathrm{P}(X_1) \\ &= -\sum_{i=1}^{d} q_i \log q_i = H_e(q), \end{aligned}$$

where $H_e(q)$ is called the entropy with respect to the probability density q_i, $i = 1, \ldots, d$.

(ii) The *average* rate of a sequence of random variables $\mathbf{Y} = \{Y_n, n \geq 1\}$, denoted by $H_{avg}(\mathbf{Y})$, is defined as follows:

$$\begin{aligned} H_{avg}(\mathbf{Y}) = \limsup_{n \to \infty} \frac{1}{n} \sum_{i_1, \ldots, i_n} (i_1 + \ldots + i_n) \\ \mathrm{P}(Y_m = i_m, m = 1, \ldots, n). \end{aligned}$$

In particular, if $\mathbf{Y}$ is a sequence of nonnegative integer-valued i.i.d. random variables with the probability density function $\mathrm{P}(Y_1 = i) = r_i, \;\; i = 0, 1, 2, \ldots$, then

$$H_{avg}(\mathbf{Y}) = E[Y_1] = \sum_{i=0}^{\infty} i r_i = H_a(r),$$

where $H_a(r)$ is called the average with respect to the probability density r_i, $i = 0, 1, 2, \ldots$

(iii) The *peak* rate of a sequence of random variables $\mathbf{Z} = \{Z_n, n \geq 1\}$, denoted by $H_{pk}(\mathbf{Z})$, is defined as follows:

$$H_{pk}(\mathbf{Z}) = sup_{n>0} ||Z_n||_\infty,$$

where $||V||_\infty$ is the essential supremum of a random variable V, i.e., $||V||_\infty = \inf\{x : \mathrm{P}(V > x) = 0\}$. In particular, if $\mathbf{Z}$ is a sequence of nonnegative integer-valued i.i.d. random variables with the probability density function $\mathrm{P}(Z_1 = i) = r_i, \;\; i = 0, 1, 2, \ldots$, then

$$H_{pk}(\mathbf{Z}) = ||Y_1||_\infty = \sup\{i : r_i > 0\} = H_p(r),$$

where $H_p(r)$ is called the peak with respect to the probability density r_i, $i = 0, 1, 2, \ldots$

In source coding, an input sequence of discrete random variables $\mathbf{X}$ on the finite alphabet $\Sigma = \{a_1, \ldots, a_d\}$ is mapped to an output sequence of nonnegative random variables $\mathbf{Y}$. The purpose of source coding is for data compression. However, there is a fundamental limit for data compression. Let Y_n be the length of the binary representation of X_n. Then, Shannon's source coding theorem (see e.g., [30, 9]) states that the average rate of $\mathbf{Y}$ must be larger than the entropy rate of $\mathbf{X}$, i.e., $H_{avg}(\mathbf{Y}) > H_{etr}(\mathbf{X})$, in order to have a decodable source code.

Analogous to source coding, a traffic regulator maps an input sequence of random variables to an output sequence of random variables. The purpose of traffic regulation is to "control" the interaction among various inputs in integrated networks so that quality of service can be guaranteed. The input of a regulator might come from the output of a source coder after a packetization process. For instance, the representation of $\mathbf{Y}$ might be changed from bits (or nats) to cells. In this paper, we shall assume the inputs of a regulator are all nonnegative integer valued. The basic unit is called a cell. In the following, we make the definition more precise.

Definition 2.2 A *regulator* is a (random) mapping that maps a nonnegative integer-valued input (arrival) process $\mathbf{Y}$ to a nonnegative integer-valued output (departure) process $\mathbf{Z} = \{Z_n, n \geq 1\}$. The *rate* of the regulator is defined to be the peak rate of $\mathbf{Z}$, i.e., $H_{pk}(\mathbf{Z})$. Let $Y(n_1, n_2) = \sum_{n=n_1+1}^{n_2} Y_n$ and $Z(n_1, n_2) = \sum_{n=n_1+1}^{n_2} Z_n$. Define the loss probability of the regulator as follows:

$$P_L = \limsup_{n \to \infty} \frac{Y(0,n) - Z(0,n)}{Y(0,n)}. \tag{1}$$

A regulator is *causal* if every cell departs after its arrival. A regulator is called *universal* if the mapping is independent of the distribution of the input sequences $\mathbf{Y}$, i.e., the mapping only depends on the sample path of the input. The rate of a universal regulator is defined to be the maximum rate among all the input sequences, i.e., $\sup_{\mathbf{Y}} H_{pk}(\mathbf{Z})$.

In the following, we show a theorem that is analogous to Shannon's source coding theorem.

Theorem 2.3 *Suppose $\mathbf{Y}$ is a stationary and ergodic sequence. Let $\mathbf{Z}$ be the output of a regulator.*

(i) *The rate of a regulator with no loss ($P_L = 0$) is not less than the average rate of $\mathbf{Y}$, i.e., $H_{avg}(\mathbf{Y}) \leq H_{pk}(\mathbf{Z})$.*

(ii) *For any R that is larger than the average rate of $\mathbf{Y}$, i.e., $R > H_{avg}(\mathbf{Y})$, there exists a* causal *regulator with a rate $H_{pk}(\mathbf{Z}) \leq R$ and an arbitrarily small loss probability.*

Proof. (i) For a regulator with no loss, one has

$$\lim_{n \to \infty} \frac{Z(0,n)}{Y(0,n)} = 1, \;\; a.s. \tag{2}$$

Since we assume $\mathbf{Y}$ is stationary and ergodic,

$$\lim_{n \to \infty} \frac{1}{n} Y(0,n) = H_{avg}(\mathbf{Y}), \;\; a.s. \tag{3}$$

In conjunction with (2) and the definition of the rate of a regulator,

$$H_{pk}(\mathbf{Z}) = \sup_{n>0} ||Z_n)||_\infty \geq \lim_{n \to \infty} \frac{1}{n} Z(0,n) = H_{avg}(\mathbf{Y}) \;\; a.s. \tag{4}$$

(ii) Consider feeding $\mathbf{Y}$ into a discrete-time queue with service rate R and buffer size x. Then the queue length at time n, denoted by q_n, is governed by the following recursive equation:

$$q_{n+1} = (\min(q_n + Y_{n+1}, x) - R)^+, \quad q_0 = 0. \qquad (5)$$

Let $\mathbf{Z}$ be the output process from such a queue. Obviously, the regulator is causal and the rate of the regulator is not larger than R since $Z(n) \le R$ for all n. Also, if $R > H_{avg}(\mathbf{Y})$, it follows from the classical queueing theory that the loss probability can be made arbitrarily small by increasing the buffer size x. ■

3 Universal block coding and block regulating for i.i.d. sources

In this section, we further explore the connections between source coding and regulating. We assume both the inputs of the coders and the regulators are i.i.d. random variables.

3.1 Universal block coding

A *fixed rate block code* of rate R for the i.i.d. sequence $\mathbf{X}$ with block size n consists of two mappings:

1. the encoder

$$f_n : \Sigma^n \mapsto \{1, 2, \ldots, e^{nR}\},$$

2. the decoder

$$g_n : \{1, 2, \ldots, e^{nR}\} \mapsto \Sigma^n.$$

Assume X_1 has the probability density function $P(X_1 = x_i) = q_i, \quad i = 1, \ldots, d$. Let $M_1(\Sigma)$ be the set of probability density functions on Σ. The probability of a decoder error for the code is

$$P_E^{(n)} = P(g_n(f_n(X_1, \ldots, X_n)) \ne (X_1, \ldots, X_n)).$$

If both the encoder and decoder are independent of the distribution of the input, the code is called *universal.*

Theorem 3.1 *For any R such that $0 \le R \le \log d$, there exists a sequence of universal codes with rate not larger than R such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log P_E^{(n)} \le - \inf_{p:H_e(p)\ge R} D(p||q), \qquad (6)$$

where $D(p||q) = \sum_{i=1}^d p_i \log \frac{p_i}{q_i}$ is the relative entropy (or the Kullback-Leibler distance) between q and p. If, furthermore, $R > H_e(q)$, then $\lim_{n\to 0} P_E^{(n)} \to 0$ exponentially fast.

Theorem 3.1 is known as the universal coding theorem (see e.g. [9, 3]). To provide the analogy between source coding and regulating, we outline the proof for Theorem 3.1 in [3]. Define the empirical distribution

$$L_n(x) = \frac{1}{n} \sum_{j=1}^n \delta_{x_j},$$

where δ_α is the unit mass at α and $x = (x_1, x_2, \ldots, x_n)$ is a typical sourceword. Note that L_n is a d-dimensional random vector and for each x, $L_n(x)$ is a d-vector in $M_1(\Sigma)$. Let $\mathcal{E}_n = \{x : H_e(L_n(x)) < R\}$. Consider the coding scheme that we assign every sourceword x in $\mathcal{E}_n$ by a unique codeword and assign one more codeword for the rest of sourcewords. The total number of codewords is thus $M_n = |\mathcal{E}_n| + 1$.

To find $|\mathcal{E}_n|$, consider the special case that $P(X_1 = x_i) = q_i = 1/d$, $i = 1, \ldots, d$. Thus, $|\mathcal{E}_n| = d^n P_{\mathbf{X}}(\mathcal{E}_n)$. From Sanov's theorem, it then follows that

$$\limsup_{n\to\infty} \frac{1}{n} P_{\mathbf{X}}(\mathcal{E}_n) \le - \inf_{p:H_e(p)<R} D(p||q), \qquad (7)$$

when $q_i = 1/d$, $i = 1, \ldots, d$. Note that

$$\inf_{p:H_e(p)<R} D(p||q) = \inf_{p:H_e(p)<R} \sum_{i=1}^d p_i \log \frac{p_i}{1/d}$$
$$= \log d - \sup_{p:H_e(p)<R} H_e(p) \ge \log d - R. \qquad (8)$$

In conjunction with (7), one has

$$\limsup_{n\to\infty} \frac{1}{n} \log M_n \le R. \qquad (9)$$

This shows the rate of the coding scheme is not larger than R.

To find $P_E^{(n)}$, note that a decoding error occurs when $x \notin \mathcal{E}_n$. Thus, $P_E^{(n)} \le P_{\mathbf{X}}(\mathcal{E}_n^c)$ and it follows Sanov's theorem again that

$$\limsup_{n\to\infty} \frac{1}{n} \log P_E^{(n)} \le - \inf_{p:H_e(p)\ge R} D(p||q). \qquad (10)$$

The infimum in (10) is larger than 0 if $R > H_e(q)$. This, in turn, implies $P_E^{(n)}$ converges to 0 exponentially fast.

3.2 Universal block regulators

Now consider feeding $\mathbf{Y}$ into a regulator. Assume $\mathbf{Y}$ is a sequence of i.i.d. random variables with the probability density function $P(Y_1 = i) = r_i$, $i = 0, 1, \ldots$. A *block regulator* for the sequence $\mathbf{Y}$ with block size n consists of a mapping $f_n : \mathcal{Z}_+^n \mapsto \mathcal{Z}_+^n$, where $\mathcal{Z}_+$ is the set of nonnegative integers. Let $y = (y_1, \ldots, y_n)$ and $f_n(y) = (f_{n,1}(y), \ldots, f_{n,n}(y))$. As in block coding, we assume that the regulator only takes the input in the block. This implies $\sum_{\ell=1}^n f_{n,\ell}(y) \le \sum_{\ell=1}^n y_\ell$. The probability that at least a loss occurs in a block for the block regulator is

$$P_L^{(n)} = P(\sum_{\ell=1}^n f_{n,\ell}(Y_1, \ldots, Y_n) < \sum_{\ell=1}^n Y_\ell). \qquad (11)$$

Note that $P_L^{(n)}$ is not the loss probability in (1) since there is probably more than one loss in a block.

Theorem 3.2 *For any $R \ge 0$, there exists a sequence of universal block regulators with rate not larger than R such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log P_L^{(n)} \le - \inf_{p:H_a(p)\ge R} D(p||r). \qquad (12)$$

If, furthermore, $R > H_a(r)$, then $\lim_{n\to 0} P_L^{(n)} \to 0$ exponentially fast.

Proof. The proof for Theorem 3.2 is parallel to that in Theorem 3.1. Define the empirical distribution

$$L_n(y) = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j},$$

where δ_α is the unit mass at α and $y = (y_1, y_2, \ldots, y_n)$ is a typical input. Let $\mathcal{E}_n = \{y : H_a(L_n(y)) < R\}$. Note that $H_a(L_n(y)) = \frac{1}{n}\sum_{\ell=1}^{n} y_\ell$. Consider the mapping $f_n(y)$ such that $f_n(y) = (R, \ldots, R)$ if $y \notin \mathcal{E}_n$. On the other hand, if $y \in \mathcal{E}_n$, then $\sum_{\ell=1}^{n} y_\ell < nR$. Let $j^* = \inf\{j : \sum_{\ell=1}^{n} y_\ell \geq jR\}$. Assign $f_{n,\ell}(y) = R$ for $\ell = 1, \ldots, j^* - 1$, $f_{n,j^*}(y) = \sum_{\ell=1}^{n} y_\ell - (j^* - 1)R$, and $f_{n,\ell}(y) = 0$, $\ell = j^* + 1, \ldots, n$. From the definition of the mapping, it is clear that the rate of f_n is not larger than R. Also, losses only occur when $y \notin \mathcal{E}_n$. Thus, $P_L^{(n)} \leq \mathrm{P}_{\mathbf{Y}}(\mathcal{E}_n^c)$ and it follows from Sanov's theorem again that

$$\limsup_{n\to\infty} \frac{1}{n}\log P_L^{(n)} \leq -\inf_{p:H_a(p)\geq R} D(p||r). \tag{13}$$

The infimum in (13) is larger than 0 if $R > H_a(r)$. This, in turn, implies $P_L^{(n)}$ converges to 0 exponentially fast. ■

Note that the universal block regulator in the proof of Theorem 3.2 may not be *causal*. To implement a causal block regulator, one can collect cells in a block and then send them in the next block. The delay in such a block regulator is thus always bounded by twice of the block size.

There is also a converse theorem to the universal coding theorem (Theorem 3.1) that shows it is the best rate available. In the following, we show the converse theorem to Theorem 3.2.

Theorem 3.3 *For any sequence of (not necessarily universal) block regulators with rate not larger than R for some $R > H_{avg}(\mathbf{Y})$,*

$$\liminf_{n\to\infty} \frac{1}{n}\log P_L^{(n)} \geq -\inf_{p:H_a(p)> R} D(p||r). \tag{14}$$

Note that the lower bound in (14) coincides with the upper bound in (12) if the function $g(R) = \inf_{p:H_a(p)\geq R} D(p||r) < \infty$ is *continuous* at R. However, this may not be the case. Consider the case that $\mathbf{Y}$ is a constant sequence with $Y_n = R$ for all n. Then $\inf_{p:H_a(p)\geq R} D(p||r) = 0$, but $\inf_{p:H_a(p)> R} D(p||r) = \infty$. In this case, the block regulator with $f_n(y) = (R, \ldots, R)$ for all y yields no loss at all.

Proof. Consider the set $\mathcal{B}_n = \{y : H_a(L_n(y)) > R\}$. Applying Sanov's theorem (lower bound) yields

$$\liminf_{n\to\infty} \frac{1}{n}\mathrm{P}(\mathcal{B}_n) \geq -\inf_{p:H_a(p)> R} D(p||q). \tag{15}$$

Since we assume that $H_{pk}(\mathbf{Z}) \leq R$,

$$\sup_{s\geq 0} Z(s, s+n) \leq n \sup_{s>0} ||Z_s||_\infty \leq nR. \tag{16}$$

Thus, there is at least one loss for $y \in \mathcal{B}_n$, and $P_L^{(n)}$ is thus bounded below by (15). ■

Remark 3.4 The purpose for assuming i.i.d. sources is to illustrate the complete analogy between block coding and block regulating. Theorem 3.2 can be extended to stationary sources that have the following limit,

$$\lim_{n\to\infty} \frac{1}{n}\log E e^{\theta Y(0,n)} = \Lambda(\theta), \quad \theta \in I\!R \tag{17}$$

where $\Lambda(\theta)$ is a differentiable function for all $\theta \in I\!R$. As an application of the Gärtner-Ellis theorem [19, 16] one has

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{n}\log P_L^{(n)}\\ &= \limsup_{n\to\infty} \frac{1}{n}\log \mathrm{P}(Y(0,n) \geq nR) \leq -\inf_{\alpha\geq R} \Lambda^*(\alpha),\end{aligned}$$

where $\Lambda^*(\alpha) = \sup_\theta[\theta\alpha - \Lambda^*(\theta)]$ is the Legendre transform of $\Lambda(\theta)$.

4 Real-time regulators

Theorem 2.3 indicates that there exists a regulator with an arbitrarily small loss probability if the rate of the regulator is larger than the average rate of the input. However, this is at the cost of large delay. The trade-off between loss probability and delay is further depicted in Theorem 3.2 and Theorem 3.3 for block regulators. One might expect that block regulators do not achieve the best trade-off. To investigate the fundamental limit for the trade-off between delay and loss probability, we consider "real-time" regulators.

Definition 4.1 A regulator is called a *real-time* regulator with maximum delay n if it is *causal* and the delay of all the departures from the regulator is at most n (units of time). Cells that cannot meet the delay constraint are lost.

Denote by $\mathbf{F}(n, R)$ the family of real-time regulators with maximum delay n and rate R.

Our main result of this section is the following theorem.

Theorem 4.2 *Suppose the input sequence* $\mathbf{Y}$ *is* stationary *and* ergodic, *and it satisfies*

$$\lim_{n\to\infty} \frac{1}{n}\log E e^{\theta Y(0,n)} = \Lambda(\theta), \quad \theta \in I\!R \tag{18}$$

for some $\Lambda(\theta)$ that is differentiable for all $\theta \in I\!R$.

(i) *For any $R > H_{avg}(\mathbf{Y})$, there exists a sequence of universal real-time causal regulators with maximum delay n and with rate not larger than R such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log P_L^{(n)} \leq -\theta^*(R)R, \qquad (19)$$

where $\theta^*(R) = \sup\{\theta > 0 : \frac{\Lambda(\theta)}{\theta} < R\}$ (with the convention that $\theta^*(R) = 0$ if $\frac{\Lambda(\theta)}{\theta} \geq R$ for all $\theta > 0$, and $\theta^*(R) = \infty$ if $\frac{\Lambda(\theta)}{\theta} < R$ for all $\theta > 0$).

(ii) *For any $R > H_{avg}(\mathbf{Y})$ and any sequence of real-time regulators with maximum delay n and with rate not larger than R,*

$$\liminf_{n\to\infty} \frac{1}{n} \log P_L^{(n)} \geq -\theta^*(R)R. \qquad (20)$$

The function $\Lambda(\theta)/\theta$ is known as the effective bandwidth function of an input (when (18) is satisfied). Since $\Lambda(\theta)$ is convex and finite for all $\theta \in I\!R$, the function $\Lambda(\theta)/\theta$ is increasing and continuous for all $\theta > 0$. For a stationary sequence $\mathbf{Y}$, one has from Jensen's inequality that $\Lambda(\theta)/\theta \geq H_{avg}(\mathbf{Y}) = E[Y_1]$ for $\theta > 0$. Also, it is clear that $\Lambda(\theta)/\theta \leq H_{pk}(\mathbf{Y}) = ||Y_1||_\infty$ for $\theta > 0$.

The rest of this section is devoted to the proof of Theorem 4.2. Analogous to source coding, our approach is to identify the optimal regulator in $\mathbf{F}(n,R)$ and show that the optimal regulator behaves as described in Theorem 4.2.

Note that a real-time regulator with rate R is a discrete-time queueing system with service rate R. A regulator in $\mathbf{F}(n,R)$ then corresponds to a scheduling policy in such a discrete-time queueing system. Among all the scheduling policies, the Earliest Due Date (EDD) policy, defined below, is of importance.

Definition 4.3 The EDD policy is operated as follows:

1. At time 0, set current time = 0.
2. Assign every arrival a due date, $n +$ current time, when it arrives.
3. Cells that cannot meet their due dates are lost.
4. Choose R cells from those still in the queueing system (regulator) with earliest due dates and send them out. If there are less than R cells in the queueing system (regulator), send them all.
5. Update the current time by 1.

It is known that the EDD policy is optimal for various performance objectives (see e.g., [8, 7]). As expected, the EDD policy is also optimal for our problem.

Lemma 4.4 *Let $L(0,t)$ be the number of losses in $(0,t]$. Then the EDD regulator minimized $L(0,t)$ for any deterministic input sequence among all the regulators in $\mathbf{F}(n,R)$.*

Lemma 4.4, in turn, implies the EDD regulator minimizes the loss probability among all the regulators in $\mathbf{F}(n,R)$. The proof for Lemma 4.4 is standard and we omit the details. First, note that one only needs to consider work conserving policies, i.e, the policies that serve cells whenever there are cells to serve. For a working conserving, one can then invoke the pairwise interchange argument (see e.g., [8, 7]) to show the optimality of EDD since all the cells have identical processing time (one unit of time).

Since all the cells have the same maximum delay, the EDD policy is simply the First Come First Served (FCFS) policy. Under the constraints of maximum delay n and rate R, the regulator for the EDD policy can be implemented by a finite buffer queue with buffer size nR. In the following, we derive an asymptotic result for the loss probability of a discrete-time queue with finite buffer. This result is related to the recently developed theory of effective bandwidth.

Lemma 4.5 *Consider a discrete-time queue with service rate R and buffer size x. Suppose the input process $\mathbf{Y}$ satisfies the conditions in Theorem 4.2. The queue length at time n, denoted by q_n, is governed by the following recursive equation:*

$$q_{n+1} = (\min(q_n + Y_{n+1}, x) - R)^+, \quad q_0 = 0. \qquad (21)$$

Denote by $P_L(x)$ the loss probability when the buffer size is x. Then for all $R > H_{avg}(\mathbf{Y})$,

$$\limsup_{x\to\infty} \frac{1}{x} \log P_L(x) = -\theta^*(R). \qquad (22)$$

Due to page limitation, the proof of Lemma 4.5 will not be presented here.

Proof. (Theorem 4.2) (i) Consider the discrete-time queue with service rate R and buffer size x in Lemma 4.5. Now let $x = nR$. Let $\mathbf{Z}$ be the output process from such a queue. Obviously, the regulator is universal and the rate of the regulator is not larger than R. We further assume the queue is operated according to the first come first served policy. Then the delay is bounded by the ratio of the buffer size nR to the service rate R. This implies the delay of every departure is at most n. The upper bound for the loss probability in (19) follows directly from the upper bound in Lemma 4.5.

(ii) Since the EDD policy in Lemma 4.4 minimizes the loss probability among all the regulators in $\mathbf{F}(n,R)$, the lower bound then follows from the lower bound in Lemma 4.5. ■

5 Conclusions

In this paper, we illustrated the analogy between traffic regulation and source coding by formalizing the traffic regulation problem. We defined the rate of a regulator to be the peak rate of the output sequence and considered the loss probability as the performance measure. Under our formulation, we showed theorems that are analogous to Shannon's source coding theorem and the universal block coding theorem. For real-time traffic regulators that have a maximum delay constraint, we identified the best trade-off between delay and loss probability and showed it is characterized by the recently developed notion of effective bandwidth.

Possible extensions of our formalism include (i) relaxation of the rate constraint of a regulator, (ii) relaxation of the assumption of the input sequence, and (iii) the analogy between queueing theory and channel coding.

For (i), one may consider the output with the linear constraint $\sup_{s\geq 0} Z(s, s+n) \leq Rn + \sigma$ for some constants R and σ (see e.g., [10]). The rate of a regulator can still be defined to be R. Intuitively, the trade-off in Theorem 4.2 still holds and the finite constant σ does not play a role when the delay n is much larger than σ.

To relax the assumption of the input sequence, one may relax the limit in (18). For instance, recent advances in long range dependent processes [2, 27, 15] allow us to consider the following limit

$$\lim_{n\to\infty} \frac{1}{n^{2(1-H)}} \log E \exp(\theta \frac{Y(0,n)}{n^{2H-1}}) = \Lambda(\theta), \qquad (23)$$

for some $1/2 < H < 1$. The parameter H, known as the Hurst parameter, not only characterizes the long range correlation of the process, but also affects its large deviation behavior. For the long range dependent processes that satisfy (23), the lower bound in Theorem 4.2 can still be applied (with appropriate change of time scale). However, the upper bound is more involved.

Channel coding has been one of the most important parts of information theory. Finding the appropriate analogy between queueing theory and channel coding is not only of theoretical interests, but also can lead to possible improvement of network design. One may start from discrete memoryless channel and consider its connections to sample path coupling in queueing theory. Research along these lines are currently under investigation and will be reported separately.

References

[1] F. Baccelli and P. Bremaud. *Elements of Queueing Theory.* New York: Springer-Verlag, 1994.

[2] J. Beran, R. Sherman, M.S. Taqqu and W. Willinger, "Long-Range Dependence variable-bit-rate video traffic," *IEEE Trans. on Commun.*, Vol. 43, pp. 1566-1579, 1995.

[3] J. Bucklew. *Large Deviation Techniques in Decision, Simulation and Estimation.* New York, NY: J. Wiley & Sons, Inc., 1990.

[4] C.S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, Vol.39, pp. 913-931, 1994. The conference version in *IEEE Conference on Decision and Control*, Vol 1, pp. 999-1010, 1992.

[5] C.S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin, "Effective Bandwidth and Fast Simulation of ATM Intree Networks," *Performance Evaluation*, Vol. 20, pp. 45-66, 1994.

[6] C.S. Chang and J.A. Thomas, "Effective bandwidth in high speed digital networks," to appear in *IEEE JSAC*, 1995.

[7] C.S. Chang and D.D. Yao, "Rearrangement, majorization and stochastic scheduling," *Math. of Oper. Res.*, Vol. 18, pp. 658-684, 1993.

[8] R.W. Conway, W.L. Maxwell and L.W. Miller, *Theory of Scheduling.* Addison-Welsey Publishing Company, Reading, M.A., 1967.

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

[10] R.L. Cruz, "A calculus for network delay, Part I: Network elements in isolation," *IEEE Tran. Inform. Theory*, Vol. 37, pp. 114-131, 1991.

[11] G. de Veciana, C. Courcoubetis and J. Walrand, "Decoupling bandwidths for networks: a decomposition approach to resource management," Submitted to *IEEE/ACM Trans. Networking*, 1993.

[12] G. de Veciana and J. Walrand, "Effective Bandwidths: call admission, traffic policing & filtering for ATM networks," Submitted to *IEEE/ACM Trans. Networking*, 1993.

[13] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Boston: Jones and Barlett Publishers, 1992.

[14] N.G. Duffield, "Exponential bounds for queues with Markovian arrivals," *preprint*, 1993.

[15] N.G. Duffied and N. O'Connel, "Large Deviations and Overflow Probabilities for the General Single-Server Queue, with Application," to appear in *Proc. Camb. Pjil. Soc.*.

[16] R.S. Ellis, "Large deviations for a general class of random vectors, " *Ann. Probab.*, Vol. 12, pp. 1-12, 1984.

[17] R.S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics.* New York: Springer-Verlag, 1985.

[18] A.I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, Vol. 1, pp. 329-343, 1993.

[19] J. Gärtner, "On large deviations from invariant measure," *Theory Probab. Appl.*, Vol. 22, pp. 24-39, 1977.

[20] R.J. Gibbens and P.J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Systems*, Vol. 9, pp. 17-28, 1991.

[21] P.W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *J. Appl. Prob*, Vol. 31A, pp. 131-156, 1994.

[22] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, Vol. 9, pp. 968-981, 1991.

[23] Z. Liu, P. Nain and D. Towsley, "Exponential bounds with application to call admission," *preprint*, 1994.

[24] F.P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, Vol. 9, pp. 5-16, 1991.

[25] G. Kesidis, J. Walrand and C.S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, Vol. 1, pp. 424-428, 1993.

[26] R. Nagarajan, J. Kurose and D. Towsley, "Local allocation of end-to-end quality-of-service in high speed networks," *preprint.*

[27] I. Norros, "A Storage Model with Self-Similar Input," *Queueing System 16*, 387-396, 1994.

[28] R.T. Rockafellar. *Convex Analysis.* Princeton: Princeton University Press, 1970.

[29] I.N. Sanov, "On the probability of large deviations of random variables," in Russian, 1957. (English translation from *Mat. Sb. (42)) in Selected Translations in Mathematical Statistics and Probability I (1961), pp. 213-244).*

[30] C.E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, Vol. 27, (pt. I) pp. 379-423, (pt. 2) pp. 623-656, 1948.

[31] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis: Queues, Communication and Computing.* 1994.

[32] D.W. Stroock. *An Introduction to the Theory of Large Deviations.* New York: Springer-Verlag, 1984.

[33] W. Whitt, "Tail probability with statistical multiplexing and effective bandwidths in multi-class queues," *Telecommunication Systems*, Vol. 2, pp. 71-107, 1993.