Bits Through Queues

Venkat Anantharam, *Member, IEEE*, and Sergio Verdú, *Fellow, IEEE*

Abstract— **The Shannon capacity of the single-server queue is analyzed. We show that the capacity is lowest, equal to e^{-1} nats per average service time, when the service time distribution is exponential. Further, this capacity cannot be increased by feedback. For general service time distributions, upper bounds for the Shannon capacity are determined. The capacities of the telephone signaling channel and of queues with information-bearing packets are also analyzed.**

Index Terms— **Shannon theory, channel capacity, queueing theory, single-server queue, telephone-signaling channel, data networks.**

I. Introduction and Summary of Results

FREQUENTLY, information is transmitted not only in the contents of messages, but also in their timing. Pauses in spoken language or the time-intervals between electronic-mail messages convey information. An immediate question prompted by this simple observation is whether interesting channel models can be constructed incorporating timing information and, if so, what is the capacity of those channels. Naturally, interesting models arise when the output is a noisy version of the input. Among the random mechanisms that blur timing information, queueing is one of the most fundamental from both the theoretical and practical viewpoints.

Consider the following simple communication channel model: an error-free bit pipe leading to a buffer modeled by a single-server queue whose "packets" or "customers" are single bits. If the service rate is μ bits per second, common wisdom would indicate that the Shannon capacity of this communication link is μ bits per second. Indeed, the service rate of a queue is often referred to as its "capacity" [7]. As we will show in this paper, this intuition is wrong: the capacity is actually *higher* than μ bits per second. How could we possibly transmit information at a rate faster than the service rate? After all, overloading the queue with an arrival rate greater than μ will surely not achieve this seemingly impossible objective, since the queue will become unstable and its output rate will not be greater than μ. The capacity is greater than μ because information can be encoded into the epochs of arrival to the queue. In fact, we can see immediately that a queue with deterministic service time has infinite capacity: the encoder can avoid queueing by sending packets spaced farther apart than the service time; then, the decoder obtains the exact timing of the arrivals and thus an infinite amount of information per arrival. For any nondeterministic service distribution, the departure times will be the result of a random transformation of the arrival times, but the receiver can still recover the transmitted information reliably as long as it is encoded on the arrival times at a rate not exceeding the capacity of the queue. This brings us to the central problem solved in this paper. Suppose now that the "packets" or "customers" are identical and thus carry no information except in their arrival times. Every message is encoded by a different sequence of n arrival times to the queue. The decoder selects one of the possible messages upon observation of the corresponding n departure times. The rate of the code is equal to the logarithm of the number of messages divided by the average time it takes to receive all n packets. The sources of randomness in this communication channel model are the service times, which are assumed to be independent and identically distributed for each packet. Despite the simplicity and canonical nature of this channel, the derivation of its capacity is far from elementary because of its various information-theoretic challenges:

- □ Because of the queueing of packets, the channel has memory.
- □ Outputs depend on inputs in a nonlinear fashion.
- □ The channel is nonstationary because the queue is assumed initially empty.
- □ Even for the single-server queue, simple queueing-theoretic results [6] are known only when the queue is in steady state and the input is "nice." When computing capacity, we cannot constrain the encoder to choose such nice arrival processes.

Channels with point-process observations have been analyzed in the information-theoretic literature (e.g., [14]) as models of direct-detection optical communication channels. In those models, the transmitter controls the instantaneous rate of the observed point process. Therefore, they are fundamentally different from the models considered in this paper.

Another motivation (outside the domain of reliable communication) for our analysis is the recent notion of resolvability [4], defined as the maximum number of random bits per second that need be generated in order to simulate (with arbitrary accuracy) the response of a given system to any input process no matter how complex. In view of the infeasibility of analytical results for queues driven by many "real-world" inputs, simulation is often the only recourse and, thus the resolvability of the queue is a very relevant quantity. The main result of [4] is that for most channels, resolvability is equal to Shannon capacity. The single-server queue is one

Manuscript received April 12, 1994; revised December 1994. The material in this paper was presented in part at the 1994 IEEE International Symposium on Information Theory, Trondheim, Norway, June 1994.

V. Anantharam is with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720 USA.

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

Publisher Item Identifier S 0018-9448(96)00439-7.

such channel; thus our capacity results are relevant not only within the domain of reliable communication but also in the domain of efficient simulation.

For the purposes of analyzing capacity, the input process is a degree of freedom; the single-server queue is characterized by its service distribution. In queueing theory, the most tractable service distribution is the exponential (or memoryless) distribution. In the usual convention, the exponential server is denoted as $\cdot/M/1$, whereas a single-server with "generic" distribution is denoted by $\cdot/G/1$. We show in this paper that exponential service turns out to play the same role in queueing systems that Gaussian noise plays in additive noise channels: exponential service leads to closed-form results and it is the service distribution with the lowest capacity for a fixed service rate. To our knowledge, this is the first time that the $\cdot/M/1$ queue has been shown to be the worst among all $\cdot/G/1$ queues according to any criterion. This extremal property may not sound surprising, as the exponential distribution maximizes differential entropy among all nonnegative random variables with fixed mean (e.g., [2]); and so in a sense it is the "noisiest" service distribution. We will see, however, that the link between both properties is far from direct.

In Section II we show the following results on the capacity of the single-server queue:

- The capacity of the $\cdot/G/1$ queue with service rate μ is greater than or equal to e^{-1} nats (0.531 b) per average service time (or $e^{-1}\mu$ nats per second).
- Among all $\cdot/G/1$ queues with given service rate, the $\cdot/M/1$ queue has the lowest capacity, equal to e^{-1} nats per average service time.
- The capacity of the $\cdot/M/1$ queue can be achieved by a random encoding strategy which for each message places $e^{-1}\mu T$ arrivals independently and uniformly distributed on a long interval of length T.
- The capacity of the $\cdot/M/1$ queue does not increase even if the encoder has full feedback information of the output of the queue. It does not decrease even if there is an unknown number of packets in the queue initially.
- The capacity of the $\cdot/G/1$ queue with service distribution S is smaller than or equal to
$$\sup_{\lambda\le\mu}\frac{\lambda}{\mu}\sup_{\substack{X\ge 0\\ E[X]\le\frac{1}{\lambda}-\frac{1}{\mu}}} I(X;X+S)$$
bits per average service time.
- The capacity of the $\cdot/G/1$ queue is smaller than or equal to $\psi(d)$ nats per average service time, where d is the divergence between the service distribution and the exponential distribution with the same mean and where $\psi(x)=x$ if $x\ge 1$ and $\psi(x)=e^{x-1}$ if $0\le x\le 1$.
- The foregoing results hold for any nonpreemptive service discipline, e.g., first-come first-serve, last-come first-serve, etc.

One of the main practical scenarios that played a role in the inception of information theory was the evaluation of the maximum rate of information that the telephone voice channel can sustain reliably. Transmitting information as fast as possible has a clear economic incentive because phone tolls depend on call duration. The minimum cost of sending one bit of information through the channel is determined by the channel capacity. Actually, it is possible to send information at a nonzero rate through the telephone channel free of charge by using the signaling channel instead of the voice channel. This is because unanswered telephone calls are always toll-free. However, their timing can carry information. Without picking up the phone, the receiver can extract that information from the time intervals between the received phone calls and from the number of rings of each phone call. The sources of randomness are the call transit times (the time it takes for the call to reach its destination). We consider a simple model of the telephone signaling channel where, upon placing the ith call, the transmitter listens until the first ring is heard, at which time he hangs up and then, after a period of time which depends on the message being sent (and possibly on previous transit times), places the $i+1$th call. In this simple model we do not exploit of the ability to send information in the number of rings allowed for each call. Call transit times are assumed to be independent and identically distributed with a generic distribution S. The receiver selects a message based on the intervals between phone calls. This channel turns out to be equivalent to a queue with feedback and it is much easier to solve than the queue without feedback analyzed in Section II. In Section III, we show that the capacity of the telephone signaling channel is equal to
$$C_F=\sup_{\beta>0}\sup_{\substack{X\ge 0\\ E[X]\le\beta}}\frac{I(X;X+S)}{E[S]+\beta}.$$

This expression is always greater than or equal to e^{-1} nats per average transit time with equality if the call transit time is exponentially distributed.

Let us now return to the problem we considered at the beginning, where information is not only sent in the timing of packets but also in their contents. So now we face the more general case where packets sent through the queue are not identical and thus contain information. For example, we can allow each packet to be selected by the encoder from a finite alphabet and then assume that each packet is received error-free by the decoder. Some asynchronous data links [1] fall within this model. More generally, we will allow packets to be corrupted by noise in their transit through the communication link. This is modeled by a channel with given capacity C_0. In Section IV we show that the capacity of the information-bearing queue is
$$C_I=\sup_{\lambda\le\mu}C(\lambda)+\lambda C_0$$
where μ is the service rate of the queue, and $C(\lambda)$ is the capacity of the queue at output rate λ. This equation reflects an inherent tradeoff: we want to inject packets to the queue at a rate λ as close as possible to the service rate; but doing so destroys the information contained in the arrival times. If C_0 is sufficiently large, then it is not worth sacrificing input rate in order to convey information via the departure times. However, in many cases of interest, it is preferable to transmit at a fraction of the permissible rate. A lower bound to C_I is shown to be $\psi(C_0)$ nats per average service time (where ψ

was defined above). This bound is achieved with equality if the server is exponential. Regardless of the service distribution, the packet alphabet, and the value of C_0, every received packet carries at least one nat of information.

For binary-valued packets going through a noiseless bit pipe, we find that the capacity of the queue is equal to

$$2e^{-1} \log_2 e \ \mu = 1.0615\mu \ \text{b/s}$$

when the server is exponential, and is larger for any other service distribution. Thus we show that reliable information transmission through queues at rates higher than the service rate is indeed possible. Moreover, the potential for improving the delay-throughput of the queue via coding of arrival timing is even more significant.

II. Single-Server Queue

A. Preliminaries

We consider the standard work-conserving single-server queueing model. According to this model, upon the departure of a packet, the server selects a packet from among those waiting, which immediately enters service, and departs from the queue after a random service time. Service times are assumed independent and identically distributed. The capacity of the system depends on its statistical description exclusively through the conditional distribution of the departure times given the arrival times. As long as the service discipline (the rule used to select the next customer to be served) is nonpreemptive and is precluded from using potential knowledge of future service times, the conditional distribution is not affected by the service discipline. Thus the results in this section hold for any service discipline that conforms to the model just described. However, it may be helpful for the reader to focus on a First-In-First-Out queue.

In principle, the rate of a code for the transmission of information through a queueing system may be defined in different ways, depending on whether the amount of information (logarithm of the codebook size) is normalized by the (average) time that it takes to *transmit* or that it takes to *receive* a codeword. Unlike conventional communication systems, this issue arises here because of the randomness in the transit time through the queue. The sensible choice for normalizing the amount of information is the average time it takes for the codewords to be received by the decoder. Appendix I illustrates the pitfalls inherent in the alternative normalization by transmission time.

Definition 1: An (n, M, T, ε)-code for a queue consists of a codebook of M codewords, each of which is a vector of n nonnegative interarrival times $(a_1, \cdots a_n)$ such that the kth arrival occurs at time $\sum_{i=1}^{k} a_i$; a decoder which upon observation of all n departures selects the correct codeword with probability greater than $1 - \epsilon$, assuming that the queue is initially empty; and the nth departure from the queue occurs on the average (over equiprobable codewords and the queue distributions) no later than T. The rate of an (n, M, T, ε)-code is defined as $(\log M)/T$.

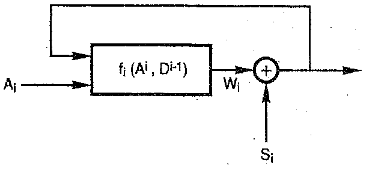

Fig. 1. Discrete-time channel model of the single-server queue.

Following the standard definitions of channel capacity (e.g., [3, p. 101] and [2, p. 194]) we define

Definition 2: The capacity C of the queue is the largest R for which for all $\gamma > 0$ there exists a sequence of (n, M, T, ε_T)-codes that satisfy

$$\frac{\log M}{T} > R - \gamma$$

and $\epsilon_T \rightarrow 0$.

Our approach to obtaining C will actually yield more information in the sense that we will obtain the capacity achievable with codes whose departure rate is fixed, defined as follows

Definition 3: R is ϵ-achievable at output rate λ if for all $\gamma > 0$ there exists a sequence of $(n, M, n/\lambda, \epsilon)$-codes such that

$$\lambda \frac{\log M}{n} > R - \gamma.$$

Rate R is achievable at output rate λ if it is ϵ-achievable at output rate λ for all $0 < \epsilon < 1$. The capacity of the queue at output rate λ, $C(\lambda)$, is the maximum achievable rate at output rate λ.

Theorem 1: The capacity of a single-server $\cdot/G/1$ queue with service rate μ satisfies

$$C = \sup_{\lambda < \mu} C(\lambda).$$

Proof: See Appendix II.

B. Converse Theorem

Theorem 2: For any $\cdot/G/1$ queue with service time S and $E[S] = 1/\mu$, if $\lambda \leq \mu$, then

$$C(\lambda) \leq \lambda \sup_{\substack{X \geq 0 \\ E[X] \leq \frac{1}{\lambda} - \frac{1}{\mu}}} I(X; X + S) \tag{2.1}$$

where X is independent of S.

Proof: The input–output map of an initially empty single-server queue is depicted in Fig. 1.

The variables A_i, D_i, S_i, and W_i denote the ith interarrival time, interdeparture time, service time, and idling time, respectively. The idling time W_i is the time elapsed between the i − 1th departure and the ith arrival; if the ith arrival occurs before the i − 1th departure, then $W_i = 0$ (cf. Fig. 2.). This means that the kth packet arrives and departs at times

$$\sum_{i=1}^{k} A_i \quad \text{and} \quad \sum_{i=1}^{k} D_i$$

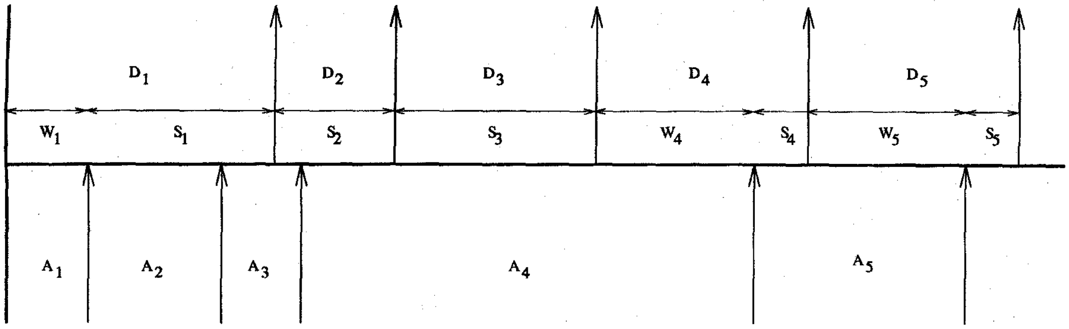

Fig. 2. Illustration of interarrival, interdeparture, service, and idling times.

respectively, and that

$$D_i = W_i + S_i. \tag{2.2}$$

Note that S_i is independent of $\{A_j\}$ and D^{i-1}. The system that generates the idling times is deterministic, causal, time-varying and has memory:

$$\begin{aligned} W_i &= f_i(A^i; D^{i-1}) \\ &= \max\left\{0, \sum_{j=1}^{i} A_j - \sum_{j=1}^{i-1} D_j\right\} \end{aligned} \tag{2.3}$$

where we have used the shorthand notation $D^i = (D_1, \cdots, D_i)$.

Denote the transmitted message by $U \in \{1, \cdots, M\}$ and the decoded message by $V \in \{1, \cdots, M\}$. Assuming that U is equiprobable, Fano's inequality [2, sec. 2.11] and the data processing lemma [2, sec. 2.8] guarantee in the usual way that every $(n, M, n/\lambda, \epsilon)$-code satisfies

$$\begin{aligned} \log M &\leq \frac{1}{1-\epsilon}[I(U;V) + \log 2] \\ &\leq \frac{1}{1-\epsilon}[I(A_1, \cdots, A_n; D_1, \cdots, D_n) + \log 2]. \end{aligned} \tag{2.4}$$

Regardless of the distributions of the arrival and service times, the n-block input–output mutual information that appears in (2.4) satisfies the following key identity:

$$\begin{aligned} I(A_1, \cdots, A_n; D_1, \cdots, D_n) &= \sum_{i=1}^{n} I(W_i; W_i + S_i) \\ &\quad - D\left(P_{D_1, \cdots, D_n} || \prod_{i=1}^{n} P_{D_i}\right). \end{aligned} \tag{2.5}$$

To check (2.5), note first that the divergence therein can be written as

$$\begin{aligned} \sum_{i=2}^{n} I(D^{i-1}; D_i) &= \sum_{i=2}^{n} I(A^n, D^{i-1}; D_i) \\ &\quad - \sum_{i=2}^{n} I(A^n; D_i | D^{i-1}). \end{aligned}$$

where the equality follows from Kolmogorov's identity:

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y).$$

By another use of this identity, we can verify that

$$I(A^n; D^n) = I(A^n; D_1) + \sum_{i=2}^{n} I(A^n; D_i | D^{i-1}).$$

Thus all that remains in order to verify (2.5) is to check that

$$I(A^n, D^{i-1}; D_i) = I(W_i; D_i). \tag{2.6}$$

But (2.6) holds because W_i is a sufficient statistic of (A^n, D^{i-1}) for D_i (which is the only property we used to obtain (2.5)).

Now define the concave monotone function

$$c(a) = \sup_{\substack{X \geq 0 \\ E[X] \leq a}} I(X; X + S) \tag{2.7}$$

where X is independent of S. Using (2.5) we obtain the bound

$$I(A_1, \cdots, A_n; D_1, \cdots, D_n) \leq \sum_{i=1}^{n} c(E[W_i]). \tag{2.8}$$

From (2.2) we see that the constraint that the expected last departure occurs before time n/λ imposes the following constraint on the average idling times:

$$\frac{1}{n}\sum_{i=1}^{n} E[W_i] = \frac{1}{n}\sum_{i=1}^{n} E[D_i] - \frac{1}{n}\sum_{i=1}^{n} E[S_i] \leq \frac{1}{\lambda} - \frac{1}{\mu}. \tag{2.9}$$

Now, uniting (2.4), (2.8), and (2.9), and using the concavity and monotonicity of the function $c(\cdot)$, we get that the rate of any $(n, M, n/\lambda, \epsilon)$-code satisfies

$$\begin{aligned} \lambda \frac{\log M}{n} &\leq \frac{\lambda}{1-\epsilon}\left(\frac{1}{n}\sum_{i=1}^{n} c(E[W_i]) + \frac{\log 2}{n}\right) \\ &\leq \frac{\lambda}{1-\epsilon}\left(c\left(\frac{1}{n}\sum_{i=1}^{n} E[W_i]\right) + \frac{\log 2}{n}\right) \\ &\leq \frac{\lambda}{1-\epsilon}\left(c\left(\frac{1}{\lambda} - \frac{1}{\mu}\right) + \frac{\log 2}{n}\right) \end{aligned}$$

which readily implies the desired result. □

In Section III, we will prove a stronger version of Theorem 2 where the capacity bound (2.1) holds even if the encoder is provided with full noiseless instantaneous feedback information of the queue output.

The two immediate questions are: 1) how to compute the maximal mutual information in the upper bound of Theorem 2, and 2) whether that upper bound is tight. We will show that if the service distribution is exponential, then bound (2.1) is tight and the maximization therein can be solved explicitly. Examining the proof of Theorem 2, we can expect that in order for the upper bound in Theorem 2 to be tight, it is enough that the inequality in (2.8) be asymptotically tight (when both sides are divided by n). For that purpose it is necessary that the service distribution be such that there exists an arrival process for which both of the following conditions hold:

- The departure process becomes a renewal process in steady state.
- Asymptotically, the idling times have first-order distributions that maximize the mutual information in (2.1).

As we will see, both conditions are satisfied in the case of the $\cdot/M/1$ queue; at present, we do not know whether there exists any other service distribution for which both of those conditions are satisfied. Note that contrary to what one might expect for the memoryless channel (2.2) there is no requirement that the idling times be independent.

Let us now turn our attention to the maximization of mutual information in (2.1). If the service time is exponential, the optimization problem is solved by the following result which is parallel to the well-known result for second-moment constrained random variables and Gaussian noise. Rather surprisingly, Theorem 3 appears to be new.

Theorem 3: Fix nonnegative scalars a and b. Let $\overline{N}$ be exponentially distributed with mean b, i.e., its pdf is

$$p_{\overline{N}}(t) = \frac{1}{b}e^{-t/b}, \quad t \geq 0.$$

Define $\overline{X}$ as a nonnegative random variable independent of $\overline{N}$ by the following mixture of a point mass and an exponential distribution:

$$P[\overline{X} = 0] = \frac{b}{a+b}$$
$$P[\overline{X} > x | \overline{X} > 0] = e^{-x/(a+b)}.$$

Note that $E[\overline{X}] = a$. Then

a)
$$I(\overline{X}; \overline{X} + \overline{N}) = \log\left(1 + \frac{a}{b}\right).$$

b) For any nonnegative random variable X (independent of $\overline{N}$) with mean a
$$I(X; X + \overline{N}) \leq I(\overline{X}; \overline{X} + \overline{N}).$$

c) For any nonnegative random variable N (possibly dependent on $\overline{X}$) with mean b
$$I(\overline{X}; \overline{X} + \overline{N}) \leq I(\overline{X}; \overline{X} + N).$$

d) For any independent nonnegative random variables X and N with means a and b, respectively
$$I(X; X + N) = \log\left(1 + \frac{a}{b}\right) + D(P_N || P_{\overline{N}}) - D\left(P_{X+N} || P_{\overline{X}+\overline{N}}\right)$$

Proof:

a) Let $\overline{Y} = \overline{X} + \overline{N}$. It is easy to verify (e.g. with Laplace transforms) that $\overline{Y}$ is exponential with mean $a + b$. Thus
$$\begin{aligned} I(\overline{X}; \overline{Y}) &= E\left[\log \frac{p_{\overline{N}}(\overline{Y} - \overline{X})}{p_{\overline{Y}}(\overline{Y})}\right] \\ &= \log\left(1 + \frac{a}{b}\right) - \frac{1}{b}E[\overline{Y} - \overline{X}] + \frac{1}{a+b}E[\overline{Y}] \\ &= \log\left(1 + \frac{a}{b}\right). \end{aligned}$$

b) For any nonnegative random variable X, independent of $\overline{N}$ with mean a
$$\begin{aligned} I(X; X + \overline{N}) &= D\left(P_{X+\overline{N}|X} || P_{X+\overline{N}} | P_X\right) \\ &= D\left(P_{X+\overline{N}|X} || Q | P_X\right) - D\left(P_{X+\overline{N}} || Q\right) \\ &\leq D\left(P_{X+\overline{N}|X} || Q | P_X\right) \qquad (2.10) \end{aligned}$$

where Q is any distribution such that $P_{X+\overline{N}} \ll Q$. In this case, we choose Q to be exponential with mean $a + b$. Then, the right-hand side of (2.10) becomes
$$\begin{aligned} D\left(P_{X+\overline{N}|X} || Q | P_X\right) &= \log\left(1 + \frac{a}{b}\right) \\ &\quad - \frac{1}{b}E[\overline{N}] + \frac{1}{a+b}E[X + \overline{N}] \\ &= \log\left(1 + \frac{a}{b}\right) \end{aligned}$$

regardless of the choice of X. This proof shows that if $P_X \neq P_{\overline{X}}$, then the inequality in b) is strict, as in that case
$$D\left(P_{X+\overline{N}} || Q\right) > 0.$$

c) Choose any nonnegative random variable N with mean b
$$\begin{aligned} I(\overline{X}; \overline{X} + N) &= D\left(P_{\overline{X}+N|\overline{X}} || P_{\overline{X}+\overline{N}|\overline{X}} | P_{\overline{X}}\right) \\ &\quad - D\left(P_{\overline{X}+N} || P_{\overline{X}+\overline{N}}\right) \\ &\quad + E\left[\log \frac{P_{\overline{Y}|\overline{X}}(\overline{X} + N | \overline{X})}{P_{\overline{Y}}(\overline{X} + N)}\right] \\ &\geq E\left[\log \frac{P_{\overline{Y}|\overline{X}}(\overline{X} + N | \overline{X})}{P_{\overline{Y}}(\overline{X} + N)}\right] \\ &= \log\left(1 + \frac{a}{b}\right) - \frac{1}{b}E[N] + \frac{1}{a+b}E[\overline{X} + N] \\ &= \log\left(1 + \frac{a}{b}\right) \end{aligned}$$

regardless of the choice of N, where the inequality follows from the fact that conditioning increases divergence [3].

d) Let $Y = X + N$. We can decompose the mutual information between X and Y as
$$\begin{aligned} I(X; Y) &= E\left[\log \frac{P_{Y|X}(Y|X)}{P_{\overline{Y}|\overline{X}}(Y|X)}\right] + E\left[\log \frac{P_{\overline{Y}|\overline{X}}(Y|X)}{P_{\overline{Y}}(Y)}\right] \\ &\quad - E\left[\log \frac{P_Y(Y)}{P_{\overline{Y}}(Y)}\right] \qquad (2.11) \end{aligned}$$

where the expectations are with respect to the joint distribution of X and Y. It is easy to see that if we condition on X in the first expectation in the right side of (2.11) we obtain the constant $D(P_N || P_{\overline{N}})$. The second expectation depends on X and Y only through their respective means, so it is equal to $I(\overline{X}; \overline{X} + \overline{N})$. Finally, the third term in (2.11) is equal to $D(P_Y || P_{\overline{Y}})$. □

A less self-contained approach to showing part b) of Theorem 3 is to use the fact that differential entropy is maximized among all nonnegative random variables with fixed mean by the exponential pdf [2, p. 269]. One can write

$$I(X; X + \overline{N}) = h(X + \overline{N}) - h(\overline{N})$$

which is maximized by making $X + \overline{N}$ exponential by the aforementioned choice of $\overline{X}$ as a mixture of a point mass at zero and a exponential distribution.

Applying the previous results to the exponential server, we obtain

Theorem 4: The $\cdot/M/1$ queue with service rate μ satisfies

$$C(\lambda) \leq \lambda \log \frac{\mu}{\lambda}, \quad \lambda \leq \mu \tag{2.12}$$

$$C \leq e^{-1}\mu \text{ nats/s.} \tag{2.13}$$

Proof: In Theorem 3, let $b = 1/\mu$ and $a = 1/\lambda - 1/\mu$. Then, Theorem 2 and Theorem 3a) and 3b) result in (2.12). To show (2.13), recall Theorem 1 and

$$\max_{\lambda \leq \mu} \lambda \log \frac{\mu}{\lambda} = e^{-1}\mu \log e \tag{2.14}$$

where the maximum is achieved at $\lambda = e^{-1}\mu$. □

In Section III, we will see that the upper bound in (2.12) holds even if the encoder has a) full output feedback information, and b) the ability to recall instantaneously the packet in service. More generally, we can get the following upper bound on the capacity of the $\cdot/G/1$ queue. This bound is the counterpart of Ihara's bound [5] (found in another form by Shannon [8]) on the capacity of the non-Gaussian noise channel. The bound depends on the shape of the service distribution only through its divergence from the exponential pdf with the same mean, which is equal to the difference between their differential entropies.

Theorem 5: Let e_μ be the exponential distribution with mean $1/\mu$. The $\cdot/G/1$ queue with service distribution S and rate $\mu = 1/E[S]$ satisfies

$$C(\lambda) \leq \lambda \log \frac{\mu}{\lambda} + \lambda D(P_S || e_\mu), \lambda \leq \mu \tag{2.15}$$

$$C \leq \begin{cases} \mu e^{-1} \exp\left(D(P_S || e_\mu)\right), & \\ & \text{if } 0 \leq D(P_S || e_\mu) \leq 1 \text{ nat} \\ \mu D(P_S || e_\mu), & \text{if } D(P_S || e_\mu) \geq 1 \text{ nat} \end{cases} \tag{2.16}$$

Proof: The upper bound in (2.15) follows immediately from (2.1) and Theorem 3d) upon dropping the last term therein. Maximizing the bound in (2.15) over all $\lambda < \mu$ yields the bound in (2.16). □

As a simple exercise, the reader may apply Theorem 5 to the uniformly distributed service time to obtain an upper bound to its capacity of 1/2 nat per average service time. On the other hand, it is easy to see that the capacity is lower bounded by 1/2 bit per average service time: at those time epochs that are multiples of $2/\mu$ we send $0/1$ packets; this ensures no queueing and the noiseless reception of the transmitted codeword. But this already belongs to the next subsection.

C. Achievability Theorem for $\cdot/M/1$ Queue

Prior to showing that the upper bounds on capacity found in Section II-B are tight, it may contribute to the reader's intuition to consider the operation of the maximum-likelihood decoder when the service distribution is exponential, and contrast it to the familiar minimum Euclidean/Hamming distance maximum-likelihood decoding in conventional channels. For convenience, let us assume that the queue is initially empty. For every codeword in the codebook, the maximum-likelihood decoder computes the service times that result in the received departure times (via (2.2) and (2.3)). If some of those service times are negative, that means that codeword could not have been transmitted. Among the codewords whose corresponding service times are nonnegative, the maximum-likelihood decoder selects the one with the smallest *sum* of service times.

The upper bounds in Theorem 4 hold with equality as Theorem 6 below demonstrates. In the proof of Theorem 6, we will use the intuition gained in the proof of Theorem 2, where we saw that for any input process to the $\cdot/G/1$ queue, the input–output mutual information satisfies

$$I(A_1, \cdots A_n; D_1, \cdots D_n) = \sum_{i=1}^{n} I(W_i; W_i + S_i) - D\left(P_{D_1 \cdots D_n} || \prod_{i=1}^{n} P_{D_i}\right).$$

Note that the last term can be thought of as the information penalty incurred by the inability of the encoder to directly control the idling times of the queue (cf. Fig. 1). For the $\cdot/M/1$-queue we can simultaneously maximize $I(W_i; W_i + S_i)$ and make

$$D\left(P_{D_1 \cdots D_n} || \prod_{i=1}^{n} P_{D_i}\right)$$

equal to zero, by letting the queue be initially in equilibrium and by choosing Poisson inputs.

Actually, in the proof of the direct theorem we do not simply maximize the input–output mutual information rate, as no previous result guarantees that the capacity of a queue is lower-bounded by its maximal input–output mutual information rate. Instead we will take advantage of the general results in [11], by means of which we need to show that the probability that the normalized input–output information density is lower than the desired bound vanishes.

Theorem 6: The $\cdot/M/1$ queue with service rate μ satisfies

$$C(\lambda) \geq \lambda \log \frac{\mu}{\lambda}, \quad \lambda < \mu \tag{2.17}$$

$$C \geq e^{-1}\mu \text{ nats/s.} \tag{2.18}$$

Proof: First note that (2.18) follows from (2.17) as in the proof of Theorem 4.

We will show that the rate in (2.17) is achievable even if the queue is initially in equilibrium rather than empty, as we have assumed heretofore. Assuming that the queue is in equilibrium allows the use of classical steady-state results.

Fix $\lambda < \mu$, let $\rho = \lambda/\mu$, and assume without loss of generality a First-In-First-Out service discipline. To start transmission, the encoder injects the so-called packet $zero$ which finds N dummy packets in the queue. The random variable N is geometrically distributed $P[N = k] = (1-\rho)\rho^k$ and is independent of the message and the transmitted codeword. Furthermore, the remaining service time of the customer in service (if any) is exponentially distributed with mean $1/\mu$. Packet $zero$ can be thought of as a synchronizing packet sent by the encoder at a time known to the decoder *a priori* (which, therefore, contains no information). The departure time of packet $zero$ is assumed to be observable by the decoder (by being informed of N or if packet $zero$ carries a marker). As we will see, by knowing the system time of packet $zero$ the decoder acquires all the initial randomness which is relevant to subsequent (message-dependent) departure times.

After packet $zero$, the encoder sends n packets whose arrival times contain all the information present in the message. We will refer to those packets as codeword packets. As before, $(A_1, \cdots, A_n)$ denote the interarrival times in the sense that the ith codeword packet ($i = 1, \cdots n$) arrives at time $\sum_{j=1}^{i} A_j$. The time elapsed between the departures of the ith and $i-1$th codeword packets is D_i. We will label the system time (queueing plus service) of packet *zero* by D_0. Note that D_0 is *not* an interdeparture time; it can be written as

$$D_0 = S_0 + U \tag{2.19}$$

where U denotes the unfinished work of the queue in equilibrium. Thus U is the sum of a geometrically distributed number of independent exponential random variables. It is easy to check that D_0 is exponentially distributed with mean $1/(\mu - \lambda)$.

The dependence of D_i on $A_1, \cdots, A_i$ and $D_0, \cdots, D_{i-1}$ is as in (2.2) and (2.3), namely

$$D_i = S_i + W_i \tag{2.20}$$

$$W_i = \max\left\{0, \sum_{j=1}^{i} A_j - \sum_{j=0}^{i-1} D_j\right\}. \tag{2.21}$$

Because of the randomness in the initial state, reliable transmission in this situation implies reliable transmission when the encoder knows that the queue is initially empty. To see this formally, let us first extend the previous code definition to an (n, M, T, ε) -code for the queue in equilibrium, where T is an upper bound to the expected exit time of the nth codeword packet (i.e., the nth packet after packet *zero*). Note that for every such code there exists an $\left(n+1, M, T, \frac{\epsilon}{1-\rho}\right)$-code for the initially empty queue. Simply use the same codebook for the empty queue where the first arrival for every codeword occurs at time $t = 0$; first, it is obvious that the expected last departure cannot increase if the queue is initially empty rather than in equilibrium; second, the probability of error of the decoder for the queue in equilibrium can be lower bounded by $(1-\rho)$ times the probability of error of the decoder for the initially empty queue because when the first codeword packet arrives to the queue in equilibrium, it finds it empty with probability $(1-\rho)$.

The only observables with which the decoder is allowed to work are $D^n = (D_0, D_1, \cdots, D_n)$. Note that D_0 is independent of the message, and can be viewed as a random initial condition of the system.

Rather than proving first that the queue is well-behaved in the sense that its capacity is lower-bounded by the limit of the maximal normalized input–output mutual information, it is preferable to take the shortcut suggested by [11]. According to those results, it is enough to show that for some input process, the liminf in probability (so-called *inf-information rate*) of the normalized information density

$$\frac{1}{n} i_{A^n;D^n}(A^n; D^n) = \frac{1}{n} \log \frac{P_{D^n|A^n}(D^n|A^n)}{P_{D^n}(D^n)} \tag{2.22}$$

is greater than or equal to $\log(\mu/\lambda)$. In other words, there exists a random sequence $A_1, A_2, \cdots$ for which

$$P\left[\frac{1}{n} i_{A^n;D^n}(A^n; D^n) < \log \frac{\mu}{\lambda} - \gamma\right]$$

vanishes for every $\gamma > 0$ as $n \to \infty$.

As we anticipated before, our choice of the input random process is Poisson with rate λ, i.e., the interarrival times $(A_1, \cdots, A_n)$ are independent exponentially distributed with mean $1/\lambda$. Since packet *zero* found the queue in equilibrium and subsequent arrivals are Poisson, Burke's $/M/M1$ output theorem [1] implies that the n interdeparture times following the departure of packet *zero* are independent exponentially distributed with mean $1/\lambda$

$$P_{D_1,\cdots,D_n}(d_1, \cdots d_n) = \prod_{i=1}^{n} e_\lambda(d_i). \tag{2.23}$$

Using the queue equations and the fact that the input process is independent of previous events in the queue, we can establish immediately that

$$P_{D^n|A^n}(d_0, \cdots, d_n|a_1, \cdots, a_n) = e_{\mu-\lambda}(d_0) \prod_{i=1}^{n} e_\mu(d_i - w_i) \tag{2.24}$$

with

$$w_i = \max\left\{0, \sum_{j=1}^{i} a_j - \sum_{j=0}^{i-1} d_j\right\}.$$

Combining (2.22), (2.23), and (2.24), we can write for any A^n and D^nconnected through the queue equations

$$\begin{aligned}\frac{1}{n} i_{A^n;D^n}(a^n;d^n) &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{e_\mu(d_i - w_i)}{e_\lambda(d_i)} \\ &\quad - \frac{1}{n} \log \frac{P_{D_0|D_1,\cdots,D_n}(d_0|d_1,\cdots,d_n)}{P_{D_0}(d_0)} \\ &= \log \frac{\mu}{\lambda} + \frac{\lambda-\mu}{n}\sum_{i=1}^{n} d_i + \frac{\mu}{n}\sum_{i=1}^{n} w_i \\ &\quad - \frac{1}{n} i_{D_0;D_1,\cdots,D_n}(d_0;d_1,\cdots,d_n). \qquad (2.25)\end{aligned}$$

We see from (2.25) that in order to prove that for all $\gamma > 0$ it suffices to show that the random sequences

a) $\displaystyle \frac{\lambda-\mu}{n}\sum_{i=1}^{n} D_i + \frac{\mu}{n}\sum_{i=1}^{n} W_i$

b) $\displaystyle \frac{1}{n} i_{D_0;D_1,\cdots,D_n}(D_0;D_1,\cdots,D_n)$

converge to 0 in probability.

Using (2.20) we can write the first random sequence as

$$\frac{\lambda-\mu}{n}\sum_{i=1}^{n} D_i + \frac{\mu}{n}\sum_{i=1}^{n} W_i = \frac{\lambda}{n}\sum_{i=1}^{n} D_i - \frac{\mu}{n}\sum_{i=1}^{n} S_i$$

which converges to 0 in probability because of the weak law of large numbers and the fact that $(D_1,\cdots,D_n)$ and $(S_1,\cdots,S_n)$ are both independent identically distributed with respective means $1/\lambda$ and $1/\mu$. To conclude the proof of Theorem 6 we will show the following result on the asymptotic degree of dependence of D_0 and $(D_1,\cdots,D_n)$.

Lemma 1: Suppose that packet *zero* arrives at a $\cdot/M/1$ queue in equilibrium and subsequent packets arrive at the times of a Poisson process whose rate is smaller than the service rate. Let D_0 be the system time of packet *zero* and let $D_1,\cdots,D_n$ be the interdeparture times following the departure of packet *zero.* Then

$$\frac{1}{n} i_{D_0;D_1,\cdots D_n}(D_0;D_1,\cdots D_n) \to 0 \text{ in probability.}$$

Proof of Lemma 1: Let Z_0 denote the random number of packets in the queue at the time of departure of packet *zero.* Since D_0 is the time that packet *zero* spends in the system, we have

$$P_{Z_0|D_0}(k|d_0) = \frac{(\lambda d_0)^k}{k!} e^{-\lambda d_0}$$

which can be averaged with respect to the unconditional distribution of D_0 (exponential with mean $(\mu-\lambda)^{-1}$, to obtain

$$P_{Z_0}(k) = (1-\rho)\rho^k$$

which is just a consequence of the well-known fact that packet *zero* leaves behind a queue in equilibrium. Then, we can write the information density as (see the bottom of this page) where in the first equality we have used the fact that D_0 and $D_1,\cdots,D_n$ are conditionally independent given Z_0 because of the memorylessness of the interarrival times and their independence from the service times; and in the second equality we used the explicit form of the conditional and unconditional distribution of Z_0.

The liminf in probability of a normalized information density cannot be negative according to [11, Theorem 8c]. Thus all we need to show is that for any $\gamma > 0$, the probability of the following event vanishes

$$\frac{\sum_{k=0}^{\infty} \frac{(\lambda D_0)^k}{k!} e^{-\lambda D_0} P_{D_1,\cdots,D_n|Z_0}(D_1,\cdots,D_n|k)}{\sum_{k=0}^{\infty} \rho^k(1-\rho) P_{D_1,\cdots D_n|Z_0}(D_1,\cdots,D_n|k)} > \exp(n\gamma).$$

Given $\eta > 0$, we can choose T_0 so large that $P[D_0 > T_0] < \eta$. On the event $\{D_0 \le T_0\}$, we can choose a large enough constant H such that

$$\frac{(\lambda D_0)^k}{k!} e^{-\lambda D_0} \le H\rho^k(1-\rho) \quad \text{for all } k \ge 0$$

and the proofs of Lemma 1 and Theorem 6 are complete. □

The proof of Theorem 6 reveals that our initial assumption that the queue is empty is not crucial for the validity of the main result that the capacity of the exponential server queue is e^{-1} nats per average service time. As the proof of Theorem 7 will demonstrate, we can even assume an arbitrary distribution for the initial condition without changing the capacity. This is not surprising in view of well-known results for compound channels [13].

The proof of Theorem 6 also reveals that a capacity-achieving random encoding strategy is to assign codewords that are independent realizations of a Poisson process with rate $e^{-1}\mu$. This is equivalent to generating for every codeword a Poisson random variable N with mean $e^{-1}\mu T$, and then distributing N arrivals uniformly and independently on an interval of length T. It is possible to simplify this type of random codebook by noticing that the entropy of N grows logarithmically with T. Therefore, we do not decrease the rate of information transmission by assuming that N is deterministic and equal to its former mean $e^{-1}\mu T$.

$$\begin{aligned} i_{D_0;D_1,\cdots,D_n}(d_0;d_1,\cdots,d_n) &= \log\left(E\left[P_{D_1,\cdots,D_n|Z_0,D_0}(d_1,\cdots,d_n|Z_0,d_0)\right]\right) - \log\left(E\left[P_{D_1,\cdots,D_n|Z_0}(d_1,\cdots,d_n|Z_0)\right]\right) \\ &= \log\left(\sum_{k=0}^{\infty} P_{D_1,\cdots,D_n|Z_0}(d_1,\cdots,d_n|k) P_{Z_0|D_0}(k|d_0)\right) - \log\left(\sum_{k=0}^{\infty} P_{D_1,\cdots,D_n|Z_0}(d_1,\cdots,d_n|k) P_{Z_0}(k)\right) \\ &= \log\left(\sum_{k=0}^{\infty} \frac{(\lambda d_0)^k}{k!} e^{-\lambda d_0} P_{D_1,\cdots,D_n|Z_0}(d_1,\cdots d_n|k)\right) - \log\left(\sum_{k=0}^{\infty} \rho^k(1-\rho) P_{D_1,\cdots,D_n|Z_0}(d_1,\cdots,d_n|k)\right). \end{aligned}$$

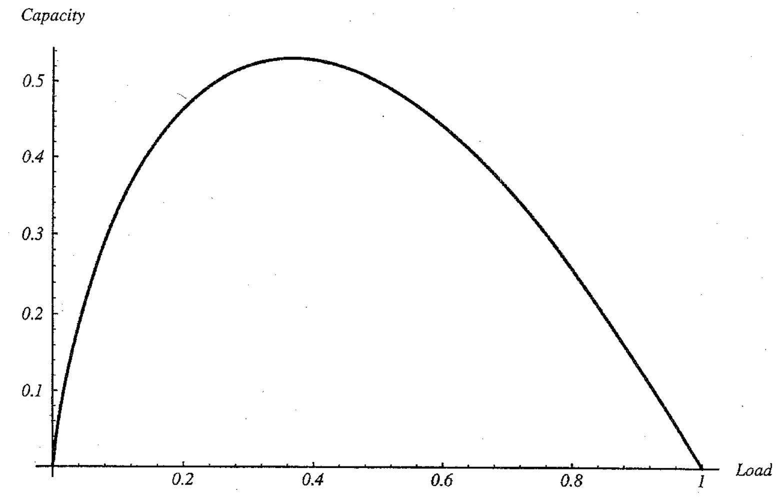

Fig. 3. Capacity (in bits per average service time) of the $\cdot/M/1$ queue as a function of the load factor λ/μ.

Theorems 4 and 6 result in the formula

$$C(\lambda) = \lambda \log \frac{\mu}{\lambda}, \quad 0 \leq \lambda \leq \mu$$

for the capacity of the $\cdot/M/1$ queue at rate λ.

This formula (plotted in Fig. 3) exhibits the precise way in which the arrival rate affects the ability to transmit information reliably. Low λ avoids queueing and turns the channel into, essentially, an additive noise channel, at the expense of a high percentage of time in which the channel is, in effect, idle. Conversely, if λ is close to μ, then long queues are the norm, the server is rarely idle, and, consequently, the departure times carry almost no information about the input timing. The optimum input rate that achieves (2.14) is $\lambda = e^{-1}\mu$.

D. Achievability Theorem for $\cdot/G/1$ Queue

We will now generalize Theorem 6 to any service distribution.

Theorem 7: The $\cdot/G/1$ queue with service rate μ satisfies

$$C(\lambda) \geq \lambda \log \frac{\mu}{\lambda}, \quad \lambda < \mu \tag{2.26}$$

$$C \geq e^{-1}\mu \quad \text{nats/s.} \tag{2.27}$$

Proof: We will proceed as in the proof of Theorem 6, assuming that packet *zero* finds the queue in equilibrium with respect to a rate-λ Poisson input process. This process need not maximize mutual information for the $\cdot/G/1$ queue, but unlike Theorem 6, the lower bound in Theorem 7 is not necessarily tight.

By means of a change of measure technique, this proof builds upon the proof of Theorem 6. We will denote the conditional and unconditional output distributions of the $\cdot/G/1$ server driven by that process by $Q_{D^n|A^n}$ and Q_{D^n}, respectively. We will retain the corresponding notation (used in the proof of Theorem 6) $P_{D^n|A^n}$ and P_{D^n}, for the case of the $M/M/1$ queue in equilibrium.

We can write the present information density as

$$\begin{aligned} \frac{1}{n} i_{A^n D^n}(a^n, d^n) &= \frac{1}{n} \log \frac{Q_{D^n|A^n}(d^n|a^n)}{Q_{D^n}(d^n)} \\ &= \frac{1}{n} \log \frac{Q_{D^n|A^n}(d^n|a^n)}{P_{D^n|A^n}(d^n|a^n)} \\ &\quad - \frac{1}{n} \log \frac{Q_{D^n}(d^n)}{P_{D^n}(d^n)} \\ &\quad + \frac{1}{n} \log \frac{P_{D^n|A^n}(d^n|a^n)}{P_{D^n}(d^n)}. \end{aligned} \tag{2.28}$$

We shall deal first with the third term in (2.28). Then, we will complete the proof by dealing with the difference between the first two terms. We saw in the proof of Theorem 6 that the third term in (2.28) satisfies

$$\begin{aligned} \frac{1}{n} \log \frac{P_{D^n|A^n}(d^n|a^n)}{P_{D^n}(d^n)} &= \log \frac{\mu}{\lambda} + \frac{\lambda - \mu}{n} \sum_{i=1}^{n} d_i \\ &\quad + \frac{\mu}{n} \sum_{i=1}^{n} w_i - \frac{1}{n} i_{D_0; D_1, \cdots, D_n}(d_0; d_1, \cdots, d_n). \end{aligned} \tag{2.29}$$

When the dummy variables are replaced by the corresponding random variables arising in the $M/G/1$ queue, we can still prove that the liminf in probability of the right-hand side of (2.29) is greater or equal to $\log \mu/\lambda$. As in the proof of Theorem 6, we have

$$\frac{\mu}{n} \sum_{i=1}^{n} W_i = \frac{\mu}{n} \sum_{i=1}^{n} D_i - \frac{\mu}{n} \sum_{i=1}^{n} S_i.$$

By the law of large numbers and the following lemma it follows that

$$\frac{\mu}{n} \sum_{i=1}^{n} W_i$$

converges in probability to $\mu/\lambda - 1$ and

$$\frac{\lambda - \mu}{n} \sum_{i=1}^{n} D_i$$

converges in probability to $1 - \mu/\lambda$.

Lemma 2: Consider an $M/G/1$ queue driven by a Poisson process of rate $\lambda < \mu$, and an arbitrary initial distribution (for the number of initial packets and the initial residual service time), then the departure of the nth exogenous (not initially present) packet satisfies

$$\frac{1}{n}\sum_{i=1}^{n} D_i \to \lambda^{-1} \quad \text{in probability.}$$

Proof: Packets can depart only after they arrive. Hence

$$D_0 + \sum_{i=1}^{n} D_i \geq \sum_{i=1}^{n} A_i.$$

Since D_0 is a proper random variable for any initial condition, it is easy to see from this that for any $\epsilon > 0$

$$P\left[\frac{1}{n}\sum_{i=1}^{n} D_i \leq \lambda^{-1} - \epsilon\right] \to 0.$$

On the other hand

$$P\left[\frac{1}{n}\sum_{i=1}^{n} D_i \geq \lambda^{-1} + \epsilon\right] \leq P\left[D_0 + \sum_{i=1}^{n} D_i \geq n/\lambda + n\epsilon\right] = P\left[X_{n/\lambda+n\epsilon} > N_{n/\lambda+n\epsilon} - n\right] \tag{2.30}$$

where X_t is the number of exogenous packets in the queue at time t, and N_t denotes the number of exogenous arrivals in $[0, t]$. A consequence of the ergodicity of the Markov chain associated with the $M/G/1$ queue is that X_t converges in distribution to a proper random variable regardless of the initial condition. It easily follows that the right-hand side of (2.30) vanishes for every $\epsilon > 0$. □

Now consider the following generalization of Lemma 1.

Lemma 3: For any $\cdot/G/1$ queue, with a departure at time D_0 and subsequent interdeparture times $D_1, \cdots, D_n$, that satisfy (2.19), (2.20), and (2.21), the limsup in probability of

$$\frac{1}{n} i_{D_0; D_1, \cdots, D_n}(D_0; D_1, \cdots, D_n)$$

is equal to 0, where the information density is computed with the distributions arising in the $\cdot/M/1$ queue with the same rate.

Proof: The proof of Lemma 1 holds verbatim, since the only condition on D_0 we used therein was that it be a proper random variable, which is still satisfied in the present (more general) case because of stability of the queue. □

To conclude the proof of Theorem 7 we will show that the liminf in probability of the sum of the first two terms in (2.28) is nonnegative (note that its expectation is nonnegative for every n because of the divergence data-processing theorem) (see the bottom of this page)

But for any $\delta > 0$

$$\begin{aligned} P\left[\frac{1}{n}\log\frac{Q_{A^n|D^n}(A^n|D^n)}{P_{A^n|D^n}(A^n|D^n)} \leq -\delta\right] &= \iint Q_{A^n|D^n}(x^n|y^n) Q_{D^n}(y^n) 1\{Q_{A^n|D^n}(x^n|y^n) \\ &\leq \exp(-\delta n) P_{A^n|D^n}(x^n|y^n)\} dx^n dy^n \\ &\leq \iint \exp(-\delta n) P_{A^n|D^n}(x^n|y^n) Q_{D^n}(y^n) dx^n dy^n \\ &= \exp(-\delta n) \end{aligned}$$

which vanishes for every $\delta > 0$. □

III. Telephone Signaling Channel

In this section we consider the telephone signaling channel introduced in Section I. Its capacity turns out to be closely related to the results we found in Section II for the single-server queue.

Upon placing the ith call, the transmitter listens (for S_i seconds) until the first ring is heard, at which time he hangs up, waits for W_{i+1} seconds (which depends on the message being sent and, possibly, on previous transit times) and places the $i+1$th call. We assume that the call transit times $S_1, S_2, \cdots$ are independent and identically distributed with a given distribution S. The receiver selects a message based on the intervals between phone calls; which we denote by $\{D_i\}$. Thus we obtain the following memoryless channel with feedback where W_i is the input and $\{S_i\}$ plays the role of additive noise:

$$D_i = W_i + S_i \tag{3.1}$$

where W_i is chosen with the knowledge of $D_1, \cdots, D_{i-1}$, or equivalently of $S_1, \cdots, S_{i-1}$. We will define codes and capacity analogously to Section II.

$$\begin{aligned} \frac{1}{n}\log\frac{Q_{D^n|A^n}(D^n|A^n)}{P_{D^n|A^n}(D^n|A^n)} - \frac{1}{n}\log\frac{Q_{D^n}(D^n)}{P_{D^n}(D^n)} &= \frac{1}{n}\log\frac{Q_{D^n|A^n}(D^n|A^n)}{P_{D^n|A^n}(D^n|A^n)} \frac{\prod_{i=1}^{n} e_\lambda(A_i)}{\prod_{i=1}^{n} e_\lambda(A_i)} \frac{P_{D^n}(D^n)}{Q_{D^n}(D^n)} \\ &= \frac{1}{n}\log\frac{Q_{A^n D^n}(A^n, D^n)}{P_{A^n D^n}(A^n, D^n)} \frac{P_{D^n}(D^n)}{Q_{D^n}(D^n)} \\ &= \frac{1}{n}\log\frac{Q_{A^n|D^n}(A^n|D^n)}{P_{A^n|D^n}(A^n|D^n).} \end{aligned}$$

Definition 4: An (n, M, T, ε)-feedback code for the telephone signaling channel consists of a codebook of M mappings $\{w_i = f_i^m(d_1, \cdots, d_{i-1}); i = 1, \cdots, n\}$; a decoder which upon observation of $(D_1, \cdots, D_n)$ selects the correct codeword with probability at least $1 - \epsilon$; and the codebook is such that the expected arrival time of the nth call is not larger than T. The rate of such a code is $(\log M)/T$.

Definition 5: The capacity C_F of the telephone signaling channel is the largest R for which for every $\gamma > 0$ there exists a sequence of (n, M, T, ε_T)-feedback codes that satisfy

$$\frac{\log M}{T} > R - \gamma$$

and $\epsilon_T \rightarrow 0$.

In (3.1) we have borrowed the same notation we used in Section II in order to highlight the strong relationship with the single-server queueing model. Actually, the telephone signal model with transit time S is equivalent (as far as computing capacity is concerned) to the single-server queue with service time S and with instantaneous and noiseless feedback of the queue output. To see this, notice that if the encoder for the queue receives feedback, the set of encoding strategies where no queueing is allowed incurs no loss of optimality. This is because the conditional output distribution generated by any strategy that allows queueing is the same as that of a modified strategy where the next arrival is not allowed to take place until the packet in service has departed. Once queueing is disallowed, the channel becomes the additive channel with feedback in (3.1).

Theorem 8: The capacity of the telephone signaling channel with transit time S is given by

$$C_F = \sup_{\beta > 0} \sup_{\substack{X \geq 0 \\ E[X] \leq \beta}} \frac{I(X; X+S)}{E[S] + \beta}. \tag{3.2}$$

Proof: Let $\mu = 1/E[S]$. Analogously to Definition 3, we can define $C_F(\lambda)$, the capacity of the telephone signaling channel at output rate λ. The proof of Theorem 1 holds almost verbatim in this case, so we can conclude that

$$C_F = \sup_{\lambda \leq \mu} C_F(\lambda). \tag{3.3}$$

We will show that

$$C_F(\lambda) = \lambda \sup_{\substack{X \geq 0 \\ E[X] \leq \frac{1}{\lambda} - \frac{1}{\mu}}} I(X; X+S). \tag{3.4}$$

The rationale is very simple: feedback may increase the capacity of the $\cdot/G/1$ queue, but it cannot increase [9] the capacity of the memoryless channel (3.1) even in the presence of cost constraints.

To prove the converse part, denote the transmitted message by $U \in \{1, \cdots, M\}$ and the decoded message by $V \in 1, \cdots, M\}$. Then

$$\begin{aligned} I(U; V) &\leq I(U; D_1, \cdots, D_n) \\ &= \sum_{i=1}^{n} I(U; D_i | D^{i-1}) \\ &= \sum_{i=1}^{n} I(W_i; D_i | D^{i-1}) \qquad (3.5) \\ &= \sum_{i=1}^{n} I(W_i; W_i + S_i) - I(D^{i-1}; D_i) \\ &\leq \sum_{i=1}^{n} I(W_i; W_i + S_i) \end{aligned}$$

where (3.5) follows from the fact that W_i is uniquely determined by the encoder from U and the feedback D^{i-1}, and D_i is conditionally independent of U and D^{i-1} given W_i. The remainder of the converse proof proceeds exactly as in the proof of Theorem 2.

In order to prove the direct part, we simply restrict attention to encoding strategies for the telephone signaling channel that do not use the feedback information. The constraint on the expected time of arrival of the last call translates into an input cost constraint (cf. (2.9))

$$\frac{1}{n} \sum_{i=1}^{n} E[W_i] \leq \frac{1}{\lambda} - \frac{1}{\mu}. \tag{3.6}$$

Obviously, the capacity of the additive-noise memoryless channel (3.1) allowing no feedback and subject to (3.6) is given by the right-hand side of (3.4). □

Corollary: The capacity of the telephone signaling channel is greater than or equal to e^{-1} nats per average transit time, with equality if and only if the transit-time distribution is exponential.

Proof: From Theorem 3a) and b) it is easy to check that in the exponential case, $C_F = e^{-1}/E[S]$. Furthermore, since the saddle point in Theorem 3 is unique, any other transit distribution will result in higher C_F. □

A quick experiment shows that it is possible to transmit about 0.4 b/s from Princeton, NJ, to Ithaca, NY, via MCI. Not a staggering rate, but you cannot beat the price. Our empirical observations show that the variance of the transit time is substantially lower than the variance of the exponential indicating that the actual capacity is likely to be way above that lower bound.

The results of this section allow us to conclude that feedback does not increase the capacity of an exponential-server queue. This property will not hold for any other service distribution unless the bound in Theorem 2 holds with equality, which at this point seems unlikely.

Actually, it is remarkable that the capacity of the exponential server remains e^{-1} nats per average service time even if in addition to having feedback, the encoder can recall the packet in service at any time it wishes. In the telephone signaling channel, this is equivalent to allowing the encoder to hang up before a call that has been placed reaches the destination. In order to see why this additional degree of freedom does not increase capacity, let us partition the time scale in segments of infinitesimal length dt. In each segment, the encoder has the ability to make the server idle or busy, regardless of its state in the previous interval. The decoder examines each segment to see whether or not there are is a departure from the queue. If the segments are indeed infinitesimal, there is

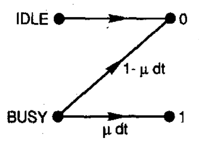

Fig. 4. Binary channel with capacity $e^{-1}\mu dt$.

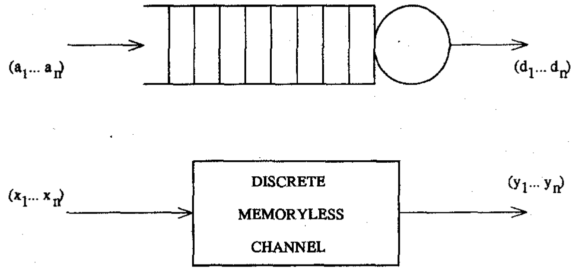

Fig. 5. Queue with information-bearing packets.

no loss of optimality in not considering the possibility of multiple departures in each segment. The channel becomes the discrete memoryless channel depicted in Fig. 4; the value of the crossover probability μdt arises because when the server is busy it puts out a Poisson stream of rate μ. It can be checked that (since dt is infinitesimal) the maximal mutual information input distribution selects BUSY with probability e^{-1}, and the capacity is equal to $e^{-1}\mu dt$ nats per infinitesimal segment. Integrating over an interval whose length is equal to the average service time we get the desired result. When both feedback and the ability to recall the packet in service are allowed, the queueing channel becomes equivalent to the on–off direct-detection photon-counting channel without dark current (see [14] for details).

IV. Queues with Information-Bearing Packets

In Section II, we studied the capacity of the single-server queue assuming that all packets were identical and thus contained no information. All the information was present in their arrival times. It is important to generalize this model to allow information-bearing packets (which may or may not be corrupted by a noisy channel). This situation is depicted in Fig. 5, and appears to be widely applicable to modeling asynchronous data link channels [1]. We will assume that the discrete-time channel in Fig. 5 is a memoryless channel (although this restriction is very easy to drop).

To fix ideas, let us consider the simplest case first: packets are binary-valued and they are received error-free. Without using any information contained in the timing of the packets we can transmit them at the rate of λ bits per second, as long as $\lambda < \mu$. Thus the capacity is at least μ bits per second. Following the results of Section II, we could send information both in the packets and in their timing. In order to transmit the information via timing, we could use an optimal coding scheme for achieving the capacity of the exponential server. As we saw in the proof of Theorem 4, the encoder would be transmitting packets at the rate of $\lambda = e^{-1}\mu$. Thus the overall information rate reliably transmitted to the decoder would be $e^{-1}\mu$ bits per second (in the binary packets) plus $e^{-1}\mu \log_2 e$ (in the timing), which is equal to 0.899μ bits per second. The reason we are doing worse than without any coding is that we have neglected the inherent tradeoff in this communication channel model: so as to transmit the maximum amount of information through the information channel, one would like to send packets at a rate arbitrarily close to the service rate of the queue; however, that would destroy any timing information present at the input. The optimum transmission rate depends, as we will see in Theorem 9, on the service distribution and on the capacity of the information channel. As we shall see, thanks to the use of timing information, the capacity of the queue with error-free transmission of binary packets is strictly greater than μ bits per second for any service distribution.

The definitions of codes and capacity in Section II are generalized in a straightforward manner to the present case. The only change is that every codeword consists now of two vectors: the vector of n interarrival times $(a_1, \cdots, a_n)$, and the vector of n symbols $(x_1, \cdots, x_n)$ drawn from alphabet A, such that the ith packet is equal to x_i and is transmitted at time $\sum_{j=1}^{i} a_j$. Throughout this section, we will assume that the queueing discipline is First-In-First-Out.

Theorem 9: The capacity of the information-bearing queue is

$$C_I = \sup_{\lambda \leq \mu} [C(\lambda) + \lambda C_0] \tag{4.1}$$

where μ is the service rate of the queue, C_0 is the capacity of the information channel (in information units per channel use), and $C(\lambda)$ is the capacity of the queue at output rate λ.

Proof: Theorem 1 can be generalized to the present problem with routine changes in its proof. Once the output rate is fixed to λ, we can view the overall channel as two independent channels in parallel (because the errors introduced by the discrete memoryless channel are assumed independent of the service times): the information channel and the queue. The capacity at output rate λ is the sum of their respective capacities: λC_0 and $C(\lambda)$. □

It should be noted that if C_0 is sufficiently large, then the λ that maximizes (4.1) is equal to μ, in which case no information is transmitted through the timing because $C(\mu) = 0$. This is evident in the following result.

Theorem 10: The capacity of the information-bearing queue with service rate μ satisfies

$$C_I \geq \begin{cases} \mu e^{-1} \exp(C_0), & \text{if } 0 \leq C_0 \leq 1 \text{ nat/symbol} \\ \mu C_0, & \text{if } C_0 \geq 1 \text{ nat/symbol} \end{cases} \tag{4.2}$$

with equality if the service time is exponential.

Proof: From Theorems 7 and 9

$$\begin{aligned} C_I &\geq \sup_{\lambda < \mu} \lambda \log \frac{\mu}{\lambda} + \lambda C_0 \qquad (4.3) \\ &= \begin{cases} \mu e^{-1} \exp(C_0), & \text{if } 0 \leq C_0 \leq 1 \text{ nat/symbol} \\ \mu C_0, & \text{if } C_0 \geq 1 \text{ nat/symbol.} \end{cases} \end{aligned}$$

Theorem 4 implies that equality is satisfied in (4.2) if the service distribution is exponential. □

The rate λ (in packets per second) that achieves the supremum in (4.3) is $\mu e^{-1}\exp(C_0)$ (which is equal to the capacity in nats per second) if this quantity is smaller than μ. Otherwise, the supremum is μC_0. This means that no matter what, we can always guarantee that each packet carries at least 1 nat of information, thanks to the addition, if necessary, of timing information. If the information channel capacity is less than 1 nat per packet and the service distribution is exponential, then that lower bound is satisfied with equality.

If we apply Theorem 10 to the important special case of a noiseless memoryless discrete channel with alphabet A, the capacity of the $\cdot/M/1$ queue with information-bearing packets is (in nats per second)

$$C_I = \begin{cases} e^{-1}\mu, & \text{if } |A| = 1 \\ 2e^{-1}\mu, & \text{if } |A| = 2 \\ \mu \log|A|, & \text{if } |A| > 2. \end{cases} \tag{4.4}$$

This provides the solution to the problem of finding the capacity of the exponential server with binary packets posed in Section I: $2e^{-1}\log_2 e\mu_0 = 1.0615\mu$ bits per second.

We conclude from (4.4) that if the channel is noiseless and the service time is exponentially distributed, timing information is worth adding only when the packets either carry no information or consist of single bits. For less "random" service distributions, including timing information is advantageous even with multibit packets. This raises the question of finding the optimum packet size in situations where there is such flexibility. In some situations it may be reasonable to assume that the service rate is inversely proportional to the packet length, i.e.

$$\mu = \frac{\mu_0}{\log|A|}.$$

In that case, (4.4) indicates that the capacity achievable with single-bit packets is $1.0615\mu_0$ versus μ_0 for multibit packets. Thus the use of single-bit packets (and timing information) increases capacity by 6% if the service time is exponentially distributed. A larger increase is possible for any other distribution. The optimal arrival rate to achieve capacity in that case is $2e^{-1}\mu = 0.74\mu$ which is roughly the target load factor in many flow control algorithms.

In contrast to Section II, the assumption that the service discipline is First-In-First-Out is crucial for the validity of the results in this section. Extending the results to other service disciplines appears to be a challenging problem.

Finally, we call attention to the fact that a potentially important application of information transmission using arrival-timing coding is in the improvement of the *delay-throughput* of the queue. Fig. 6 shows system delay D (in units of average service time) versus throughput η (in bits per average service time) for an exponential server with binary-valued packets. Three curves are shown:

- $M/M/1$ delay-throughput (upper curve) [6]:

$$D = \frac{1}{1-\eta}.$$

- $D/M/1$ delay-throughput (middle curve), where D is the root of the equation [6]

$$1 - \frac{1}{D} = e^{-\frac{1}{\eta D}}.$$

- $\cdot/M/1$ delay-throughput (lower curve) when arrivals are encoded in order to maximize throughput. From the results of this section, a Poisson input process achieves the following parametric form as a function of the input rate $\lambda \in (0, 2/e)$:

$$D(\lambda) = \frac{1}{1-\lambda}$$
$$\eta(\lambda) = \lambda + \lambda \log_2 \frac{1}{\lambda}.$$

The $D/M/1$ curve gives the best delay-throughput achievable without coding of arrival timing [12, App. B]. A remarkable improvement (for throughputs greater than 0.36 bits per average service time) is obtained once coding of arrival timing is introduced, even though the lower curve is not the optimum delay-throughput that can be obtained with arrival timing coding (as it reflects the situation in which coding is introduced so as to maximize throughput regardless of system delay). The maximum average delay incurred with the coded system is 3.78 (times the average service time) at throughput of 1.06 bits per average service time. With the same delay, the best throughput achievable with no coding is 0.86. Conversely, achieving throughput of 1 bit per average service time with an uncoded system would require unbounded delays, whereas with suitable coding of the arrival times, an average delay of twice the service time suffices to achieve the same throughput.

Appendix I

Some codewords that may take a very short time to be transmitted (by sending all the packets in a very short initial burst) take much longer to be received in their entirety. Since the system is not emptied after the transmission of that burst, transmission of the next message must wait until the whole codeword is received. If the amount of information transmitted is normalized by the time it takes to *transmit* the codeword (rather than to receive it), then, that unavoidable waiting time is not properly accounted for, and results different from those in Section II are obtained. As an illustration, compare the result we obtained on the capacity of the exponential server being equal to $e^{-1}\mu$ nats per second with the following result.

Theorem 11: An exponential server with service rate equal to μ satisfies for every n

$$\sup_{A_1,\cdots,A_n} \frac{I(A_1,\cdots,A_n; D_1,\cdots,D_n)}{E\left[\sum_{i=1}^{n} A_i\right]} = \mu \text{ nats/s.} \tag{A1.1}$$

Proof: We can use a simple extension of the results on channel capacity per unit cost [10] to simplify the optimization problem in (A1.1)

$$\sup_{A_1,\cdots,A_n} \frac{I(A_1,\cdots,A_n; D_1,\cdots,D_n)}{E[\sum_{i=1}^{n} A_i]}$$
$$= \sup_{a_1,\cdots,a_n} \frac{D\big(P_{D^n|A^n=(a_1,\cdots,a_n)} \| P_{D^n|A^n=(0,\cdots,0)}\big)}{a_1+\cdots+a_n}. \tag{A1.2}$$

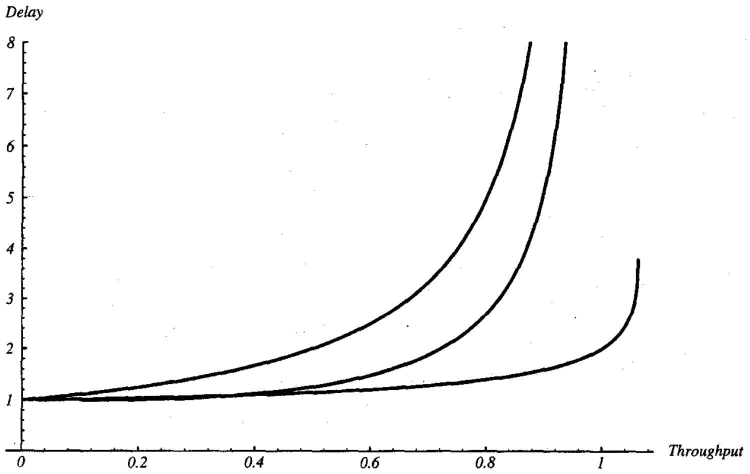

Fig. 6. Delay-throughput for exponential server and binary packets. Poisson arrivals (upper), deterministic interarrivals (middle) and no arrival-timing coding; with arrival-timing coding (lower).

The numerator in the right-hand side of (A1.2) is the unconditional divergence of the received process given the interarrival times $(a_1, \cdots, a_n)$ and $(0, \cdots, 0)$. Note that the optimization in the right-hand side of (A1.2) is over the set of n interarrival times, rather than over the set of joint nth-order distributions as in the left side.

The distribution of the departure times given that all the arrivals occur at $t = 0$ is

$$P_{D^n|A^n=(0,\cdots,0)}(d_1, \cdots, d_n) = \prod_{i=1}^{n} e_\mu(d_i). \qquad \text{(A1.3)}$$

Since the queue is assumed to be initially empty, we can write

$$P_{D^n|A^n=(a_1,\cdots,a_n)}(d_1, \cdots, d_n) = \prod_{i=1}^{n} e_\mu(d_i - w_i) \qquad \text{(A1.4)}$$

with

$$w_i = \max\left\{0, \sum_{j=1}^{i} a_j - \sum_{j=1}^{i-1} d_j\right\}.$$

Using the specific form of the exponential distribution, the divergence between the distributions in (A1.4) and (A1.3) is

$$D\left(P_{D^n|A^n=(a_1,\cdots,a_n)} \| P_{D^n|A^n=(0,\cdots,0)}\right) = \mu \sum_{i=1}^{n} E[W_i|(A_1, \cdots, A_i) = (a_1, \cdots, a_i)]. \qquad \text{(A1.5)}$$

In other words, the divergence is equal to μ times the total expected idling time prior to the last arrival given a specific set of arrival times.

Now the maximization in (A1.2) boils down to

$$\mu \sup_{a_1,\cdots,a_n} \frac{\sum_{i=1}^{n} E[W_i|(A_1, \cdots, A_i) = (a_1, \cdots, a_i)]}{a_1 + \cdots + a_n} \qquad \text{(A1.6)}$$

So we need to find the n arrival times that maximize the average total server idle time prior to the nth arrival divided by the nth arrival time. This is simple, because this ratio is obviously upper-bounded by 1 and if $(a_1, \cdots, a_n) = (a, 0, \cdots, 0)$, then the bound is satisfied with equality. □

Appendix II

Proof of Theorem 1: It is clear from the definitions that $C \geq C(\lambda)$ for all $\lambda < \mu$. To prove the desired result we will see that if there is $\alpha > 0$ such that

$$C > C(\lambda) + \alpha$$

for all $\lambda < \mu$, we will arrive at a contradiction.

Choose $\beta > 0$ and $\gamma > 0$ small enough that

$$\frac{(1-\beta)}{(1+\beta)}(C - \gamma) > C - \alpha \qquad \text{(A2.1)}$$

and

$$\frac{1}{1 + \beta\mu^{-1}}(C - \gamma) > C - \alpha. \qquad \text{(A2.2)}$$

Now select a sequence of (n, M, T, ε_T)-codes such that $\epsilon_T \to 0$ and

$$\frac{\log M}{T} > C - \gamma.$$

Denote

$$l = \limsup_{n\to\infty} n/T.$$

Note that $l \leq \mu$ because the nth departure time exceeds the sum of the service times of packets $1, \cdots, n$, whose mean is equal to $1/\mu$. We will distinguish two cases:

a) $l > 0$:

We can find a subsequence of n such that

$$(1-\beta)l < \frac{n}{T} < (1+\beta)\, l.$$

For those blocklengths, there exist $\left(n, M, \frac{n}{(1-\beta)l}, \epsilon\right)$-codes whose rate satisfies

$$\begin{aligned}(1-\beta)\, l\, \frac{\log M}{n} &> \frac{(1-\beta)}{(1+\beta)} \frac{\log M}{T} \\ &> \frac{(1-\beta)}{(1+\beta)}(C-\gamma) \\ &> C-\alpha\end{aligned}$$

contradicting the fact that

$$C > C(l(1-\beta)) + \alpha.$$

b) $l = 0$:

We now have a sequence of (n, M, T, ε_T)-codes such that $\epsilon_T \to 0$, $n/T \to 0$, and $\frac{\log M}{T} > C - \gamma$. Let $m = \beta T$. For any $\delta > 0$ and sufficiently large n we are going to derive an $\big(m, M, m\big(\beta^{-1} + \mu^{-1} + \delta\big), \epsilon_T + \epsilon_m\big)$-code, where $\epsilon_m \to 0$. The first $(m-n)$ arrivals occur at time $t = 0$. If they have not exited by time $(m-n)\big(\mu^{-1} + \delta\big)$ we declare the entire codeword in error. This happens with probability less than ϵ_m for some $\epsilon_m \to 0$. The next n arrivals are those of the original code shifted by $(m-n)\big(\mu^{-1} + \delta\big)$. Thus the expected time of the last departure is upper-bounded by $T + m\big(\mu^{-1} + \delta\big)$. The rate of the resulting $\big(m, M, m\big(\beta^{-1} + \mu^{-1} + \delta\big), \epsilon_T + \epsilon_m\big)$-code is

$$\begin{aligned}\frac{1}{\beta^{-1} + \mu^{-1} + \delta} \frac{\log M}{m} &= \frac{1}{1 + \beta(\mu^{-1} + \delta)} \frac{\log M}{T} \\ &\geq \frac{1}{1 + \beta(\mu^{-1} + \delta)}(C-\gamma) \\ &> C-\alpha\end{aligned}$$

for sufficiently small δ, because of (A2.2). But this contradicts the assumption that

$$C > C\left(\frac{1}{\beta^{-1} + \mu^{-1} + \delta}\right) + \alpha. \qquad \square$$

References

[1] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Englewood-Cliffs, NJ: Prentice-Hall, 1992.

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

[3] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[4] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. 39, pp. 752–772, May 1993.

[5] S. Ihara, "On the capacity of channels with additive non-Gaussian noise," *Inform. Contr.*, vol. 37, pp. 34–39, 1978.

[6] L. Kleinrock, *Queueing Systems. Vol. 1: Theory*. New York: Wiley, 1975.

[7] ——, "On the modeling and analysis of computer networks," *Proc. IEEE*, vol. 81, pp. 1179–1190, Aug. 1993.

[8] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July–Oct. 1948.

[9] ——, "Certain results in coding theory for noisy channels," *Inform. Contr.*, vol. 1, pp. 6–25, Sept. 1957.

[10] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1019–1030, Sept. 1990.

[11] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1147–1157, July 1994.

[12] J. Walrand, *An Introduction to Queueing Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[13] J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. New York: Springer-Verlag, 1978.

[14] A. D. Wyner, "Capacity and error exponent for the direct detection photon channel," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1449–1471, Nov. 1988.