

Bits Through Queues

Venkat Anantharam and Sergio Verdú
Cornell University Princeton University

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/sec, common wisdom would indicate that the Shannon capacity of this communication link is μ bits/sec. As we show in this paper, that intuition is wrong: the answer is actually *higher* than μ bits/sec. How could we possibly transmit information at a rate faster than the service rate? After all, overdriving the queue with an arrival rate higher than μ will not do, as the queue will become unstable and its output rate will not be higher than μ . The capacity (revealed at the end of this summary) is higher than μ because *information can be encoded into the arrival epochs*.

Let us consider now a simpler and more fundamental problem, namely the Shannon capacity of the single-server queue. Suppose that “packets” 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 channel 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 the various challenges it presents to the information theorist:

- Because of the queuing of packets, the channel has memory, and 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 queuing theoretic results exist only when the queue is in steady-state and the input is “nice”. As far as computing capacity, we cannot constrain the encoder to choose such type of arrival processes.

Since for the purposes of analyzing capacity, the input process is a degree of freedom for the encoder, the single-server queue is characterized by its service distribution. Following convention, the exponential server is denoted as $\bullet/M/1$, whereas a single-server with “generic” distribution is denoted by $\bullet/G/1$. We show in this paper that in the information theoretic analysis of queueing systems, exponential service turns out to play the same role 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 $\bullet/M/1$ queue has been shown to be the worst among all $\bullet/G/1$ queues according to any criterion.

Our main results on the capacity of the single-server queue are:

- The capacity of the $\bullet/G/1$ queue with service rate μ is greater than or equal to e^{-1} nats (0.531 bits) per average

service time (or μ/e nats per second).

- Among all $\bullet/G/1$ queues with given service rate, the $\bullet/M/1$ queue has the lowest capacity, equal to e^{-1} nats per average service time.
- The capacity of the $\bullet/M/1$ queue when the input rate is constrained to be λ is equal to $\rho \log 1/\rho$ nats per average service time, where $\rho = \lambda/\mu$.
- The capacity of the $\bullet/M/1$ 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 $\bullet/G/1$ queue with service distribution S is smaller than or equal to

$$\sup_{\lambda \leq \mu} \frac{\lambda}{\mu} \sup_{x \geq 0, E[X] \leq \frac{1}{\mu} - \frac{\lambda}{\mu}} I(X; X+S) \quad (1)$$

nats per average service time.

- The capacity of the $\bullet/G/1$ 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 $\psi(x) = x$ if $x \geq 1$ and $\psi(x) = e^{x-1}$ if $0 \leq x \leq 1$.
- The foregoing results do not depend on the service discipline, e.g. FIFO, LIFO, etc.

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. We will allow that the packets are corrupted by noise in their transit through the communication link. This is modeled by a channel with given capacity C_0 . We show that the capacity of the information-bearing queue is

$$C_I = \sup_{\lambda \leq \mu} C(\lambda) + \lambda C_0 \quad (2)$$

where $C(\lambda)$ is the capacity of the queue at output rate λ . Equation (2) reflects the tradeoff inherent in this situation. On one hand, we want to inject packets to the queue at a rate as close as possible to the service rate; on the other hand, 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 with 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; this bound is achieved with equality if the server is exponential.

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

$$2 e^{-1} \mu \log e = 1.0615 \mu \text{ bits/sec.} \quad (3)$$

if the server is exponential, or larger for any other service distribution. Thus, reliable information transmission through queues at rates higher than the service rate is indeed possible.