Qubits through Queues: The Capacity of Channels with Waiting Time Dependent Errors

Krishna Jagannathan, Avhishek Chatterjee
Department of Electrical Engineering
IIT Madras, Chennai, India

Prabha Mandayam
Department of Physics
IIT Madras, Chennai, India

***Abstract*—We consider a setting where qubits are processed sequentially, and derive fundamental limits on the rate at which classical information can be transmitted using quantum states that decohere in time. Specifically, we model the sequential processing of qubits using a single server queue, and derive explicit expressions for the capacity of such a 'queue-channel.' We also demonstrate a sweet-spot phenomenon with respect to the arrival rate to the queue, i.e., we show that there exists a value of the arrival rate of the qubits at which the rate of information transmission (in bits/sec) through the queue-channel is maximised. Next, we consider a setting where the average rate of processing qubits is fixed, and show that the capacity of the queue-channel is maximised when the processing time is deterministic. We also discuss design implications of these results on quantum information processing systems.**

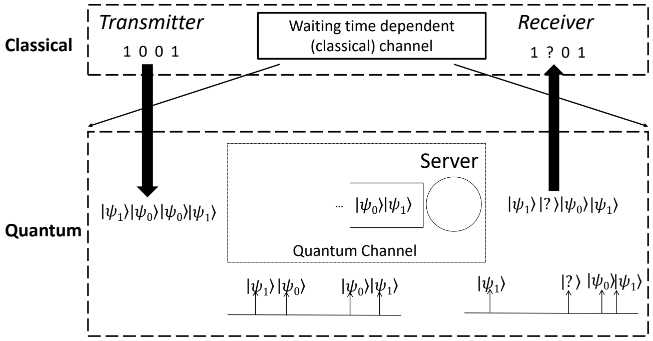
Classical
Transmitter
1 0 0 1
Waiting time dependent (classical) channel
Receiver
1 ? 0 1
Quantum
$|\psi_1\rangle|\psi_0\rangle|\psi_0\rangle|\psi_1\rangle$
Server
$\cdots\ |\psi_0\rangle|\psi_1\rangle$
Quantum Channel
$|\psi_1\rangle|?\rangle|\psi_0\rangle|\psi_1\rangle$
$|\psi_1\rangle|\psi_0\rangle$
$|\psi_0\rangle|\psi_1\rangle$
$|\psi_1\rangle$
$|?\rangle$ $|\psi_0\rangle|\psi_1\rangle$

Fig. 1. Schematic of the queue-channel depicting the case of quantum erasure.

I. Introduction

Quantum bits (or qubits) have a tendency to undergo rapid decoherence in time, due to certain fundamental physical phenomena. The manner and mechanism of such decoherence depends on the underlying physical implementation of the quantum state, the environment in which the quantum state evolves, and other physical factors such as temperature. Once a state decoheres, the information stored is lost either partially or completely, depending again on the underlying realizations and physical processes.

In this paper, we are concerned with *sequential processing of a stream of qubits* — for example, this can include transmitting, storing or performing gate operations on the quantum states. In this setting, we derive fundamental bounds on the rate at which information can be conveyed using quantum states that decohere in time.

When quantum states are prepared and then processed sequentially, it is reasonable to posit that there will inevitably be a non-zero 'processing time,' corresponding to each qubit, which in turn corresponds to a finite rate at which the qubits can be processed by the system. For example, when the qubit is prepared and transmitted as a photon polarization state, the rate at which a receiver can detect (and hence process) the photons is constrained by the average dead-time of the detectors, which is typically of the order of a tens of nanoseconds [1]. To consider another concrete example, superconducting Josephson junction based qubits have gate processing times ranging from a few tens of nanoseconds to a few hundreds of nanoseconds, while their average coherence times are typically of the order of a few tens of microseconds [2, Table 2]. In such a scenario, the coherence time of each qubit is only about two to three orders of magnitude longer than the time it takes to process each qubit. This brings us to an interesting phenomenon, which does not seem to arise naturally in transmitting classical bits: as the qubits wait to be processed, they inevitably undergo decoherence, leading to errors.

More generally, we may consider a setting where qubits are prepared (or "arrive") according to some random process at a particular rate, and are to be processed sequentially. Due to the non-zero processing time for each qubit, the arriving qubits will have to wait in sequence to be processed. The present paper focusses on obtaining a quantitative characterization of the above phenomenon. Specifically, we model the sequential processing of qubits using a single-server queue with average service rate μ. Now, suppose that the qubits 'arrive' at the queue according to a stationary random process of rate λ. Since the queue is stable if and only if $\lambda < \mu$, it is immediately clear that this system cannot process qubits at a rate higher than μ. A key question we address in this paper is as follows: Assuming for simplicity that each qubit is used to encode one classical bit, is it possible to transmit information through the above queue at a rate that is arbitrarily close to μ bits/sec?

In the case of classical bits, the answer is clearly in the affirmative. However, in the quantum case, we show that the answer turns out to be in the negative in general. Intuitively, when the arrival rate λ is very close to μ, the waiting time

978-1-5386-9286-8/19/$31.00 © 2019 IEEE

for each qubit becomes very large. As a result, most of the qubits are likely to suffer decoherence, which leads to a higher probability of error.

Indeed, under physically well-motivated models for the decoherence of qubits with time, we derive explicit expressions for the capacity of the above 'queue-channel[1].' In particular, we demonstrate a 'sweet-spot' phenomenon with respect to the arrival rate, i.e., we show that there exists a particular value of arrival rate $\lambda^* \in (0, \mu)$ at which the rate of information transmission (in bits/sec) through the queue-channel is maximised.

Next, for a given average rate μ of processing qubits, and Poisson arrivals of qubits, we prove that the channel capacity is maximised when the processing time of each qubit is deterministic. In other words, given a processing rate μ, the rate of information transmission is maximised by ensuring that the processing time is deterministic for each qubit.

Finally, we remark that similar waiting time dependent errors can also be observed in other emerging as well as classical systems. For example, due to the short-lived nature of human attention, the performance of a human deteriorates with the waiting time [4]. In this context, a waiting time dependent channel arises due to human impatience instead of quantum decoherence. This is particularly relevant to crowdsourcing. In the context of age of information [5], packets become useless (erased) after waiting in a queue for a certain duration — a scenario which also falls within the scope of the model we consider.

A. Related Literature

Gallager and Telatar initiated the area of multiple access queues in [6] which is the first published work at the intersection of queuing and information theory. Around the same time Anantharam and Verdú considered timing channels where information is encoded in the times between consecutive information packets, and these packets are subsequently processed according to some queueing discipline [7]. Due to randomness in the sojourn times of packets through servers, the encoded timing information is distorted, which the receiver must decode. In contrast to [7], *we are not concerned with information encoded in the timing between packets — in our work, all the information is in the symbols.*

An information theoretic notion of reliability of a queuing system with state-dependent errors was introduced and studied in [3], where the authors considered queue-length dependent errors motivated mainly by human computation and crowdsourcing.

B. Contributions

We focus on a simple queue-channel with waiting time induced *erasures* — specifically, we model the decoherence of qubits using a *quantum erasure channel* [8]. In the simplest $\mathsf{M/M/1}$ setting, we explicitly characterize the capacity of the erasure queue-channel, and show that there is an optimal arrival rate $\lambda_{\mathsf{M/M/1}} \in (0, \mu)$ at which the capacity of the queue-channel is maximised. Next, we generalize the above result to an $\mathsf{M/GI/1}$ setting, and show a similar behaviour.

This result highlights an unusual interplay that exists between transmission rate and delay in the quantum case. Unlike in the classical case where we can obtain any rate that is arbitrarily close to the server rate (at the expense of delay), in the quantum case, it is desirable to operate away from the server capacity from the point of view of maximising capacity. This is because when qubits are sent faster than the optimal rate, the effective rate of information transmission actually *decreases*, due to a drastic increase in waiting time induced erasures.

While the above results characterize the optimal arrival rate of qubits for a fixed service distribution, one can also ask after the best service time distribution for *fixed* values of arrival and service rates. Indeed, we show that the capacity of the queue-channel is maximised when the service time distribution is deterministic. In other words, the $\mathsf{M/D/1}$ queue maximises the queue-channel capacity, among all $\mathsf{M/GI/1}$ queues. In certain physical realizations, there could be fundamental physical constraints that translate to an optimal gate processing rate of the qubits (see for example [9]). Our result offers an important design insight in such a scenario — the capacity is maximised when the gate processing time is deterministic across qubits, i.e., it is desirable to mitigate 'jitter' in the processing times.

II. System Model

The model we study is depicted in Fig. 1. Specifically, a source generates a classical bit stream, which is encoded into qubits. These qubits are sent sequentially to a single server queue according to a stationary point process of rate λ. The server works like a FIFO queue with independent and identically distributed (i.i.d.) service times for each qubit. After getting processed by the server, each qubit is measured and interpreted as a classical bit. We refer to this system as a queue-channel, and characterize the classical capacity of this system (in bits/sec).

In order to capture the effect of decoherence due to the underlying quantum channel, we model the error probability as an explicit function of the waiting time W in the queue. For instance, in several physical scenarios, the decoherence time of a single qubit maybe modelled as an exponential random variable. In other words, the probability of a qubit error/erasure after waiting for a time W is given by $p(W) = 1 - e^{-\kappa W}$, where $1/\kappa$ is a characteristic time constant of the physical system under consideration [10, Section 8.3].

A. Queuing Discipline

We consider a continuous-time system. The service requirements are i.i.d. across qubits. The service time of the jth qubit is denoted S_j and has a cumulative distribution F_S. The average service rate of each qubit is μ, i.e., $\mathbf{E}_{F_S}[S] = 1/\mu$. In the interest of simplicity and tractability, we assume Poisson arrivals, i.e., the time between two consecutive arrivals is i.i.d. with an exponential distribution with parameter λ. For stability

[1] A terminology we borrow from [3].

of the queue, we assume $\lambda < \mu$. For ease of notation let us assume $\mu = 1$. (Our results easily extend to general μ).

Let A_j and D_j be the arrival and the departure epochs of jth qubit, respectively and $W_j = D_j - A_j - S_j$ be the time that jth qubit waits in queue until its service begins.

B. Error Model

As the qubits wait to be served, they undergo decoherence, leading to errors at the receiver. This decoherence is modelled in general as a completely positive trace preserving map [10]. However, in this paper, we restrict ourselves to a rudimentary setting where we use a fixed set of orthogonal quantum states (say $|\psi_0\rangle$ and $|\psi_1\rangle$, corresponding to classical bits 0 and 1, as depicted in Fig. 1) to encode the classical symbols at the sender's side, and measure the qubits in some *fixed* basis at the receiver's end.

In general, the jth symbol $X_j \in \mathcal{X}$, is encoded as one of a set of orthogonal states $\{|\psi_{X_j}\rangle\}$ belonging to a Hilbert space $\mathcal{H}$ of dimension $|\mathcal{X}|$. The noisy output state $|\tilde{\psi_j}\rangle$ is measured by the receiver in some fixed basis, and decoded as the output symbol $Y_j \in \mathcal{Y}$. This measurement induces a conditional probability distribution $\mathbf{P}(Y_j|X_j, W_j)$, which we can think of as an *induced classical channel* from $\mathcal{X}$ to $\mathcal{Y}$.

An n-length transmission over the waiting time dependent queue-channel is denoted as follows. Inputs are $\{X_j : 1 \leq j \leq n\}$, channel distribution $\prod_j \mathbf{P}(Y_j|X_j, W_j)$, and outputs are $\{Y_j : 1 \leq j \leq n\}$.

Throughout, $Z^k = (Z_1, Z_2, \ldots, Z_k)$ denotes a k-dimensional vector and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n, \ldots)$ denotes an infinite sequence of random variables. Information is measured in bits and $\log$ means logarithm to the base 2.

III. Queue-channel Capacity

We are interested in defining and computing the information capacity of the queue-channel, which is simply the capacity of the induced classical channel defined above. As mentioned earlier, we restrict ourselves to using a fixed set of orthogonal states to encode the classical symbols at the sender's side, and measuring in some fixed basis at the receiver's end. For this reason, the capacity of the induced classical channel is not the same as the *classical capacity of the quantum channel* resulting from the underlying decoherence model. The latter consideration is left for future work; see Section V.

A. Definitions

Let M be the message transmitted from a set $\mathcal{M}$ and $\hat{M} \in \mathcal{M}$ be its estimate at the receiver.

Definition 1: An $(n, \tilde{R}, T)$ code consists of the encoding function $X^n = f(M)$ and the decoding function $\hat{M} = g(X^n, A^n, D^n)$, where the cardinality of the message set $|\mathcal{M}| = 2^{n\tilde{R}}$, and for each codeword, the expected total time for all the symbols to reach to the receiver is less than T.

Definition 2: If the decoder chooses $\hat{M}$ with average probability of error less than ϵ, that code is said to be ϵ-achievable. For any $0 < \epsilon < 1$, if there exists an ϵ-achievable code $(n, \tilde{R}, T)$, the rate $R = \frac{\tilde{R}}{T}$ is said to be achievable.

Definition 3: The information capacity of the queue-channel is the supremum of all achievable rates for a given arrival process with distribution F_A and is denoted by $C(F_A)$ bits per unit time.

We assume that the transmitter knows the arrival process statistics, but not the realizations before it does the encoding. However, depending on the application, the receiver may or may not know the realization of the arrival and the departure time of each symbol.

Proposition 1: The capacity of the queue-channel (in bits/sec) described in Sec. II is given by

$$C(F_A) = \lambda \sup_{\mathbf{P}(\mathbf{X})} \underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y}|\mathbf{W}), \tag{1}$$

when the receiver knows the arrival and the departure time of each symbol. On the other hand, when the receiver does not have that information, the capacity is,

$$C(F_A) = \lambda \sup_{\mathbf{P}(\mathbf{X})} \underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y}). \tag{2}$$

Here, $\underline{\mathbf{I}}$ is the usual notation for inf-information rate [11].

This result is essentially a consequence of the general channel capacity expresssion in [11]. Please see [12] for details. The following lemma which is used in proving Prop. 1 would be needed later.

Lemma 1: Under the assumptions in Sec. II, $\{W_j\}$ is a Markov process and has a unique limiting distribution π.

Proof: Follows from the stability results for $\mathsf{GI/GI/1}$. ■

B. Remarks

Before proceeding further, we note the difference between the maximum symbol throughput (number of symbols processed per unit time) and the maximum information throughput (our notion of capacity) of the queuing system studied here. The symbol throughput is the maximum rate of arrivals for which the queue is stable and hence, increases with λ on $[0, \mu)$. On the other hand, the expression for 'information throughput' has λ as a multiplicative factor. However, this does not mean it increases with λ. In typical queuing systems, the average waiting time is increasing in λ. For quantum channels and other systems like crowd-sourcing and multimedia communication, service errors are more likely when waiting times are larger. Thus, increasing λ also negatively impacts the inf-information term in the capacity expression. Hence there is typically an information throughput-optimal $\lambda \in (0, \mu)$. This will be clear when we discuss some particular scenarios of interest.

The capacity expression in Prop. 1 does not provide clear insights into the behaviour of the system under different arrival and service statistics. Our main contribution in this paper lies in deriving a single letter capacity expression for queue-channels with waiting time induced *erasures*. A single letter capacity expression facilitates a clearer understanding of the effects of the arrival and service processes on the capacity. Analogous capacity results can also be derived for a class of M-ary symmetric queue-channels [12], but we omit discussing them here due to page constraints.

IV. Erasure Queue-Channels

Erasure channels are ubiquitous in classical as well as quantum information theory. We consider a quantum erasure channel [8] which acts on the jth state $|\psi_{X_j}\rangle$ as follows: $|\psi_{X_j}\rangle$ remains unaffected with probability $1-p(W_j)$, and is erased to a state $|?\rangle$ with probability $p(W_j)$, where $p : [0,\infty) \to [0,1]$ is typically increasing. Such a model also captures the communication scenarios where information packets become useless (erased) after a deadline. For such an erasure channel, a single letter expression for capacity can be obtained.

Theorem 1: For the erasure queue-channel defined above, the capacity is $\lambda \ \log|\mathcal{X}| \ \mathbf{E}_\pi \left[1 - p(W)\right]$ bits/sec, irrespective of the receiver's knowledge of the arrival and the departure times of symbols.

Proof: The proof uses an upper-bound on $\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y})$ in terms of unconditional sup-entropy rate and conditional inf-entropy rate and shows that the capacity expression is an upper-bound. On the other hand, using a similar lower-bound on $\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y})$ we show that for a choice of distribution of $\{X_n\}$ (namely, i.i.d. uniform), $\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y})$ is no smaller than the capacity expression. The fact that in the case of erasure channels the received symbol is either correct or erased (but never wrong) makes the knowledge of the arrival and departure times irrelevant. Finally, ergodicity of the queue is used in reducing n-symbol bounds for $\underline{\mathbf{I}}$ to single-letter bounds.

More precisely, using the properties of limit superior and limit inferior, we have

$$\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y}|\mathbf{W}) \leq \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{W}) - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X},\mathbf{W}),$$

where $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{W})$ and $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X},\mathbf{W})$ are the lim-sup in probability of the sequences $\frac{1}{n}\log\frac{1}{\mathbf{P}(Y^n|W^n)}$ and $\frac{1}{n}\log\frac{1}{\mathbf{P}(Y^n|X^n,W^n)}$, respectively. By the channel model, given W_i and X_i, Y_i is independent of any other variable and hence, a product form would emerge. Also, note that $\mathbf{P}(Y_i|X_i,W_i) = p(W_i)$ if Y_i is an erasure, else, it is $1-p(W_i)$. Combining these observations we obtain

$$\frac{1}{n}\log\frac{1}{\mathbf{P}(Y^n|X^n,W^n)} = -\frac{1}{n}\sum_{i=1}^{n}\big[\mathbf{1}(Y_i=\mathcal{E})\log(1-p(W_i)) + \mathbf{1}(Y_i\neq\mathcal{E})\log(p(W_i))\big], \tag{3}$$

where $\mathcal{E}$ represents erasure. By Lemma 1 the limit of the expression in (3) exists almost surely as a finite constant. Hence, this limit is the value of $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X},\mathbf{W})$. Note that for an erasure queue-channel this limit does not depend on the distribution of X^n.

Let us now consider upper-bounding $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{W})$. Let $N^{\mathcal{E}}$ be the set of indices for which $Y_i = \mathcal{E}$. Following standard conditional probability arguments, we get

$$\begin{aligned}&\log\frac{1}{\mathbf{P}(Y^n|W^n)}\\ &= -\sum_{i=1}^{n}\left[\mathbf{1}(Y_i=\mathcal{E})\log(1-p(W_i)) + \mathbf{1}(Y_i\neq\mathcal{E})\log(p(W_i))\right]\\ &\quad -\log\mathbf{P}(\{Y_i : i\notin N^{\mathcal{E}}\}|\{Y_i\neq\mathcal{E} : i\notin N^{\mathcal{E}}\},\{W_i : i\notin N^{\mathcal{E}}\})\,.\end{aligned} \tag{4}$$

Please see [12] for the detailed steps. As the almost sure limit of $\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}(Y_i=\mathcal{E})\log(1-p(W_i))\right.$ $\left.+\ \mathbf{1}(Y_i\neq\mathcal{E})\log(p(W_i))\right]$ is equal to $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X},\mathbf{W})$, we only need to focus on the second term in (4).

Note that for an erasure channel, if Y_i is not an erasure, Y_i has the same value as that of X_i. So, for any joint distribution $\mathbf{P}_{\mathbf{X}}$ of input symbols:

$$\begin{aligned}&\frac{1}{n}\ \big(-\log\mathbf{P}(\{Y_i : i\notin N^{\mathcal{E}}\}|\{Y_i\neq\mathcal{E} : i\notin N^{\mathcal{E}}\},\{W_i : i\notin N^{\mathcal{E}}\})\ \big)\\ &= -\frac{1}{n}\log\mathbf{P}_{\mathbf{X}}(\{Y_i : i\notin N^{\mathcal{E}}\}|\{W_i : i\notin N^{\mathcal{E}}\})\\ &= -\frac{n-|N^{\mathcal{E}}|}{n}\frac{1}{n-|N^{\mathcal{E}}|}\log\mathbf{P}_{\mathbf{X}}(\{Y_i : i\notin N^{\mathcal{E}}\})\end{aligned} \tag{5}$$

Note that in the limit, by Lemma 1, $\frac{|N^{\mathcal{E}}|}{n}$ converges almost surely to $\mathbf{E}[p(W)] < 1$. So, almost surely $n - |N^{\mathcal{E}}| \to \infty$. So as $n\to\infty$, $\frac{1}{n-|N^{\mathcal{E}}|}\log\mathbf{P}_{\mathbf{X}}(\{Y_i : i\notin N^{\mathcal{E}}\})$ can be seen as $\frac{1}{N}\log\mathbf{P}(X^N)$ for some large N. lim-sup in probability of this quantity is upper-bounded by $\log|\mathcal{X}|$. Thus we get an upper-bound of $(1-\mathbf{E}[p(W)])\log|\mathcal{X}|$ on $\underline{I}$.

To obtain a lower bound, we have

$$\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y}|\mathbf{W}) \geq \underline{\mathbf{H}}(\mathbf{Y}|\mathbf{W}) - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X},\mathbf{W}), \tag{6}$$

where $\underline{\mathbf{H}}(\mathbf{Y}|\mathbf{W})$ is the lim-inf in probability of $\frac{1}{n}\log\frac{1}{\mathbf{P}(Y^n|W^n)}$. We have already derived an expression for the second term above. Then, using (4) and similar arguments which lead to (5) we obtain that the right hand side of (6) equals (5).

If we choose uniform and i.i.d $\{X_i\}$, (5) almost surely converges to $(1-\mathbf{E}[p(W)])\log|\mathcal{X}|$. Thus, we derived a $\mathbf{P}_{\mathbf{X}}$ independent upper-bound on $\underline{I}(\mathbf{X};\mathbf{Y}|\mathbf{W})$ which is matched by a particular choice of $\mathbf{P}_{\mathbf{X}}$. Thus, for the erasure channel

$$\sup_{\mathbf{P}_X}\underline{I}(\mathbf{X};\mathbf{Y}|\mathbf{W}) = (1-\mathbf{E}[p(W)])\log|\mathcal{X}|.$$

By multiplying with λ we obtain the capacity of this channel. See [12] for a complete proof. ■

This single letter capacity expression allows us to mine deeper insights on system design. It is well known in queuing that waiting time increases with increasing arrival rate. As $p(\cdot)$ is increasing, so is $\mathbf{E}_\pi\left[p(W)\right]$ in λ. Therefore, it is apparent from the single letter expression (in Theorem 1) that capacity may not be monotonic in λ. This raises an interesting question: is there an optimal λ at which the capacity is maximised? The answer to this question depends on the queuing dynamics. Therefore, we first attempt to understand it for the most fundamental queuing system in communication networks, the M/M/1 queue. Interestingly, for the M/M/1 queue, there exists a simple characterization of the capacity and the corresponding optimal arrival rate.

Theorem 2: The arrival rate that maximises the information capacity of the M/M/1 queue-channel is given by

$$1 - \arg\min_{u\in(0,1)} u\ \left(1+\tilde{p}\left(\frac{u}{1-u}\right)\right) \tag{7}$$

where for any $u > 0$, $\tilde{p}(u) := \int \exp(-ux)p(x)dx$ is the Laplace transform of $p(\cdot)$.

Proof: This proof uses the exponential waiting time distribution of $\mathsf{M/M/1}$ queue to relate the capacity to Laplace transform of $p(\cdot)$.

It is known that the waiting time in $\mathsf{M/M/1}$ is distributed as $\exp\left(\frac{1-\lambda}{\lambda}\right)$ for $\mu = 1$. Thus,

$$\mathbf{E}[p(W)] = \int_0^\infty p(w)\frac{1-\lambda}{\lambda}\exp\left(\frac{1-\lambda}{\lambda}w\right)dw$$
$$= \frac{1-\lambda}{\lambda}\tilde{p}\left(\frac{1-\lambda}{\lambda}\right).$$

Thus, the capacity is given by $\lambda\left(1 - \frac{1-\lambda}{\lambda}\tilde{p}\left(\frac{1-\lambda}{\lambda}\right)\right)$.

So, the capacity maximising arrival rate is the one that maximises this expression:

$$\arg\max_{\lambda\in(0,1)} \lambda\left(1 - \frac{1-\lambda}{\lambda}\tilde{p}\left(\frac{1-\lambda}{\lambda}\right)\right)$$
$$\iff \arg\max_{\lambda\in(0,1)} \left(\lambda - (1-\lambda)\tilde{p}\left(\frac{1-\lambda}{\lambda}\right)\right)$$
$$\iff 1 - \arg\max_{u\in(0,1)} \left(1 - u - u\tilde{p}\left(\frac{u}{1-u}\right)\right)$$
$$\iff 1 - \arg\min_{u\in(0,1)} u\ \left(1 + \tilde{p}\left(\frac{u}{1-u}\right)\right).$$

■

In the case of quantum erasure channels [8], decoherence of qubits over time gives rise to an interesting form for $p(\cdot)$, namely, $p(W) = 1 - \exp(-\kappa W)$, where κ is a physical parameter. A detailed quantum physical discussion on this can be found in [13]. A relation of this kind between waiting time and erasure is also relevant in multimedia communication with deadlines and in the context of age of information. In these scenarios, when deadlines or maximum tolerable age of information packets are unknown, the exponential distribution (being the most entropic) serves as a reasonably good stochastic model. Such a model is captured by the above form of $p(\cdot)$. Hence, for this particular form of $p(\cdot)$, it is important to understand the capacity behaviour explicitly.

Corollary 1: For an $\mathsf{M/M/1}$ erasure queue-channel with $p(W) = 1 - \exp(-\kappa W)$, $F_A(x) = 1 - \exp(-\lambda\ x)$ and $F_S(x) = 1 - \exp(-x)$,

(i) the capacity is given by $\frac{\lambda(1-\lambda)}{1-\alpha\lambda}$ bits/sec, and
(ii) the capacity is maximised at

$$\lambda_{\mathsf{M/M/1}} = \frac{1}{\alpha}\left(1 - \sqrt{1-\alpha}\right) = \frac{1}{1+\sqrt{1-\alpha}}$$

where $\alpha = \frac{1}{1+\kappa}$.

This result offers interesting insights into the relation between the information capacity and the characteristic time-constant of the quantum medium. In the case of decohering channels, a larger value of the decoherence exponent κ corresponds to a faster decoherence. We note that α decreases as κ increases and hence, $\lambda_{\mathsf{M/M/1}}$ decreases as κ increases. This implies that when the qubits decohere more rapidly, the arrival rate that maximises the capacity is lower. In other words, when the coherence time is small, it is better to send at a slower rate to avoid excessive waiting time induced errors.

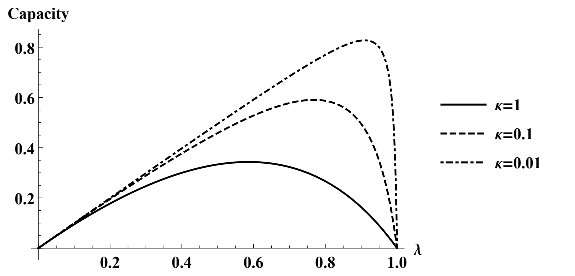

Fig. 2. The capacity of the $\mathsf{M/M/1}$ queue-channel (in bits/sec) plotted as a function of the arrival rate λ for different values of the decoherence parameter κ.

Fig. 2 depicts a capacity plot of the $\mathsf{M/M/1}$ queue-channel (in bits/sec), as a function of the arrival rate λ for different values of the decoherence parameter κ. Since the service rate μ is taken to be unity, we note that a value of $\kappa = 0.01$ corresponds to an average coherence time which is two orders of magnitude longer than the service time — a setting reminiscent of superconducting qubits [2]. We also notice from the shape of the capacity curve for $\kappa = 0.01$ that there is a drastic drop in the capacity, if the system is operated beyond the optimal arrival rate $\lambda_{\mathsf{M/M/1}}$. This is due to the drastic increase in delay induced decoherence as the arrival rate of qubits approaches the server capacity.

Next, we discuss the generalization of Corollary 1 to $\mathsf{M/GI/1}$ queues. Specifically, a result similar to Corollary 1 also holds for $\mathsf{M/GI/1}$ system for a different α, though unlike for the $\mathsf{M/M/1}$ queue, the waiting time is not exponentially distributed.

Theorem 3: For an $\mathsf{M/GI/1}$ erasure queue-channel with $p(W) = 1 - \exp(-\kappa W)$, $F_A(x) = 1 - \exp(-\lambda\ x)$, and a general F_S with $F_S(0) = 0$

(i) the capacity is given by $\frac{\lambda(1-\lambda)}{1-\alpha\lambda}$ bits/sec, and
(ii) the capacity is maximised at

$$\lambda_{\mathsf{M/GI/1}} = \frac{1}{\alpha}\left(1 - \sqrt{1-\alpha}\right) = \frac{1}{1+\sqrt{1-\alpha}},$$

where $\alpha = \frac{1-\tilde{F}_S(\kappa)}{\kappa}$, and $\tilde{F}_S(u) = \int \exp(-ux)dF_S(x)$ is the Laplace transform of the service time distribution.

Proof: First, note that for the particular form of $p(\cdot)$ considered here, the capacity expression in Theorem 1 simplifies to $\lambda\mathbf{E}[\exp(-\kappa W)]$.

Next, using the Pollaczek-Khinchin formula for the $\mathsf{M/GI/1}$ queue (with $\mu = 1$), we write

$$\mathbf{E}[\exp(-\kappa W)] = \frac{(1-\lambda)\kappa}{\kappa - \lambda(1-\tilde{F}_S(\kappa))} = \frac{1-\lambda}{1-\alpha\lambda},$$

where $\alpha = \frac{(1-\tilde{F}_S(\kappa))}{\kappa}$. Thus, the capacity is $\frac{\lambda(1-\lambda)}{1-\alpha\lambda}$ bits/sec, and the capacity maximising arrival rate is $\arg\max_{\lambda\in[0,1)} \frac{\lambda(1-\lambda)}{1-\alpha\lambda}$.

The objective function in the above optimization problem is concave in λ. This implies that the value of λ that maximises the capacity is the one at which the derivative of the capacity with respect to λ is zero. Taking the derivative, we obtain a quadratic function in λ which when equated to zero yields two solutions for λ: $\frac{1}{\alpha} \pm \frac{\sqrt{1-\alpha}}{\alpha}$. The only valid solution for which $\lambda \in [0, 1)$ is given by $\frac{1}{\alpha} - \frac{\sqrt{1-\alpha}}{\alpha}$. ■

The above results characterize an optimal λ for given arrival and service distributions. One can also ask after the best service distribution for a given arrival process and a fixed server rate. This question is of interest in designing the server characteristics like gate operations [9] or photon detectors in the case of quantum systems, as well as in the case of packet communication with age of information constraints. The following theorem is useful in such scenarios.

Theorem 4: For an erasure queue-channel with $p(W) = 1 - \exp(-\kappa W)$ and $F_A(x) = 1 - \exp(-\lambda\ x)$ at any λ the capacity is maximised by $F_S(x) = \mathbf{1}(x \geq 1)$, i.e., a deterministic service time maximises capacity, among all service distributions with unit mean and $F_S(0) = 0$.

Proof: As derived in the proof of Theorem 3, the capacity is

$$\frac{\lambda(1-\lambda)\kappa}{\kappa - \lambda(1 - \tilde{F}_S(\kappa))} = \frac{\frac{(1-\lambda)\kappa}{\lambda}}{\frac{\kappa-\lambda}{\lambda} + \tilde{F}_S(\kappa)}.$$

Thus, for any given λ, among all service distribution with unit mean, the capacity is maximised by that service distribution for which $\tilde{F}_S(\kappa)$ is minimised. For any service random variable S, by Jensen's inequality, we have $\tilde{F}_S(\kappa) = \mathbf{E}[\exp(-\kappa S)] \geq \exp(-\kappa \mathbf{E}[S])$. Therefore, $\tilde{F}_S(\kappa)$ is minimised by $S = \mathbf{E}[S]$, i.e., a deterministic service time. ■

Although we have considered only the erasure queue-channel in this paper, analogous results can be obtained for a more general class of channels which include binary (and M-ary) symmetric queue-channels. Please see [12] for details.

V. Concluding Remarks and Future Work

In this paper, we used simple queue-channel models to characterize the capacity of channels with waiting time dependent errors. Though our main motivation stems from quantum communications, where we characterize the rate at which classical information can be transmitted using orthogonal quantum states that decohere in time, the model also captures scenarios in crowdsourcing and multimedia streaming.

We believe there is ample scope for further work along several directions. Firstly, it is important to move away from the restriction of using only orthogonal states at the encoder and a fixed measurement at the receiver, and allow for arbitrary superposition states at the encoder, and arbitrary measurements at the receiver. This would allow us to invoke the true classical capacity of the underlying (non-stationary) quantum channel, in terms of a quantity [14] analogous to the classical inf-information rate. It remains an interesting technical challenge to obtain a formula for the queue-channel capacity in this general scenario, and identify channels for which the classical coding strategy would still be optimal. Furthermore, we can also consider other widely studied quantum channel models, such as the phase damping and amplitude damping channels.

We have only considered uncoded quantum bits in this paper. We can also quantitatively evaluate the impact of using quantum codes to protect qubits from errors. Employing a code would enhance robustness to errors, but would also increase the waiting time due to the increased number of qubits to be processed. It would be interesting to characterise this tradeoff, and identify the regimes where using coded qubits would be beneficial or otherwise.

More broadly, we believe our work highlights the importance of explicitly modelling delay induced errors in quantum communications. As quantum computing takes strides towards becoming an ubiquitous reality, we believe it is imperative to develop processor architectures and algorithms that are informed by more quantitative studies of the impact of delay induced errors on quantum information processing systems.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments. Thanks also to Uday Khankhoje and Shweta Agrawal for that serendipitous lunch conversation which led to the model considered here.

References

[1] R. H. Hadfield, "Single-photon detectors for optical quantum information applications," *Nature photonics*, vol. 3, no. 12, p. 696, 2009.

[2] G. Wendin, "Quantum information processing with superconducting circuits: a review," *Reports on Progress in Physics*, vol. 80, no. 10, p. 106001, 2017.

[3] A. Chatterjee, D. Seo, and L. R. Varshney, "Capacity of systems with queue-length dependent service quality," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3950 – 3963, Jun. 2017.

[4] D. Kahneman, *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[5] M. Costa, M. Codreanu, and A. Ephremides, "Age of information with packet management," in *Proc. 2014 IEEE Int. Symp. Inf. Theory*, Jun. 2014.

[6] İ. E. Telatar and R. G. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 963–969, Aug. 1995.

[7] V. Anantharam and S. Verdú, "Bits through queues," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 4–18, Jan. 1996.

[8] C. H. Bennett, D. P. DiVincenzo, and J. A. Smolin, "Capacities of quantum erasure channels," *Physical Review Letters*, vol. 78, no. 16, p. 3217, 1997.

[9] C. Ballance, T. Harty, N. Linke, M. Sepiol, and D. Lucas, "High-fidelity quantum logic gates using trapped-ion hyperfine qubits," *Physical review letters*, vol. 117, no. 6, p. 060504, 2016.

[10] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[11] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[12] A. Chatterjee, K. Jagannathan, and P. Mandayam, "Qubits through queues: The capacity of channels with waiting time dependent errors," arXiv:1804.00906, 2018.

[13] M. Grassl, T. Beth, and T. Pellizzari, "Codes for the quantum erasure channel," *Physical Review A*, vol. 56, no. 1, p. 33, 1997.

[14] M. Hayashi and H. Nagaoka, "General formulas for capacity of classical-quantum channels," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1753–1768, 2003.