A Quantum Analog of Huffman Coding

Samuel L. Braunstein*, Christopher A. Fuchs†,
Daniel Gottesman‡, Hoi-Kwong Lo
Networked Systems Department
HP Laboratories Bristol
HPL-98-23
February, 1998

E-mail: schmuel@sees.bangor.ac.uk
cfuchs@cco.caltech.edu
gottensma@t6-serv.lanl.gov
hkl@hplb.hpl.hp.com

Huffman coding, variable length codes, instantaneous codes, quantum information, quantum computing

We analyse a generalization of Huffman coding to the quantum case. In particular, we notice various difficulties in using instantaneous codes for quantum communication. However, for the *storage* of quantum information, we have succeeded in constructing a Huffman-coding inspired quantum scheme. The number of computational steps in the encoding and decoding processes of N quantum signals can be made to be polynomial in log N by a massively parallel implementation of a quantum gate array. This is to be compared with the O(N^3) computational steps required in the sequential implementation by Cleve and DiVincenzo to the well-known quantum noiseless block coding scheme by Schumacher. The powers and limitations in using this scheme in communications are also discussed.

Internal Accession Date Only
*SEECS, University of Wales, Bangor, United Kingdom
†Norman Bridge Laboratory of Physics, California Institute of Technology, Pasadena, California
‡Los Alamos National Laboratory, Los Alamos, New Mexico

A quantum analog of Huffman coding

Samuel L. Braunstein,[a*] Christopher A. Fuchs[b†],
Daniel Gottesman,[c‡], and Hoi-Kwong Lo[d§]
[a]*SEECS, University of Wales, Bangor LL57 1UT, UK*
[b]*Norman Bridge Laboratory of Physics 12-33,*
California Institute of Technology, Pasadena, CA 91125, USA
[c] *T-6 Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[d]*Hewlett-Packard Labs, Filton Road, Stoke Gifford, Bristol BS12 6QZ, UK*
(January 27, 1998)

Abstract

We analyse a generalization of Huffman coding to the quantum case. In particular, we notice various difficulties in using instantaneous codes for quantum communication. However, for the *storage* of quantum information, we have succeeded in constructing a Huffman-coding inspired quantum scheme. The number of computational steps in the encoding and decoding processes of N quantum signals can be made to be polynomial in $\log N$ by a massively parallel implementation of a quantum gate array. This is to be compared with the $O(N^3)$ computational steps required in the sequential implementation by Cleve and DiVincenzo of the well-known quantum noiseless block coding scheme by Schumacher. The powers and limitations in using this scheme in communication are also discussed.

I. INTRODUCTION

This paper is an attempt to find a source coding scheme analogous to Huffman coding in the classical source coding theory. Let us recapitulate the result in a classical theory. Consider the simple example of a memoryless source that emits a sequence of independent, identically distributed signals each of which is chosen from a list $w_1, w_2, \cdots, w_n$ with probabilities $p_1, p_2, \cdots, p_n$. The task of source coding is to store such signals with a minimal

*schmuel@sees.bangor.ac.uk

†cfuchs@cco.caltech.edu

‡gottesma@t6-serv.lanl.gov

§hkl@hplb.hpl.hp.com

amount of resources. In classical information theory, resources are measured in bits. The standard coding scheme to use is the Huffman coding algorithm. Apart from being highly efficient, it has the advantage of being instantaneous, i.e., unlike block coding schemes, the encoding and decoding of each signal can be done immediately. Note also that codewords of variable lengths are used to achieve efficiency. As we will see below, these two features—instantaneousness and variable length—of Huffman coding are difficult to generalize to the quantum case.

In the *quantum* case, we are given a quantum source which emits a time sequence of independent identically distributed pure state quantum signals each of which is chosen from $|u_1\rangle, |u_2\rangle, \cdots |u_m\rangle$ with probabilities $q_1, q_2 \cdots, q_m$ respectively. Notice that $|u_i\rangle$'s are normalized (i.e., unit vectors) but not necessarily orthogonal to each other.[1] The goal of quantum source coding is to minimize the number of dimensions of the Hilbert space needed for the faithful noiseless encoding of quantum signals. It is convenient to measure the dimensionality of a Hilbert space in terms of the number of qubits (i.e., quantum bits).

II. DIFFICULTIES IN A QUANTUM GENERALIZATION

To illustrate the difficulties involved, we shall first attempt a naive generalization of Huffman coding to the quantum case.[2] Consider the density matrix for each signal $\rho = \sum q_j |u_j\rangle\langle u_j|$ and diagonalize it into

$$\rho = \sum_i p_i |\phi_i\rangle\langle\phi_i| \,, \tag{1}$$

where $|\phi_i\rangle$ is an eigenstate and the eigenvalues p_i's are arranged in decreasing order. Huffman coding of a corresponding classical source with the same probability distribution p_i's allows one to construct a one-to-one correspondence between Huffman codewords h_i and the eigenstates $|\phi_i\rangle$. Any input quantum state $|u_j\rangle$ may now be written as a sum over the complete set $|\phi_i\rangle$. Remarkably, this means that the length of each signal is a quantum mechanical variable with its value in a superposition of the length eigenstates. It is not clear what this really means nor how to deal with such an object. If one performs a measurement on the length variable, irreversible changes to the N signals will be introduced which disastrously reduce the *fidelity*. For a pure input state $|a_i\rangle$, the fidelity of the output B_i is defined as the probability for it to pass a yes/no test of being the state $|a_i\rangle$. Mathematically, it is given by $\langle a_i|B_i|a_i\rangle$ [6].

Therefore, to faithfully encode the signals, the sender and the receiver are forbidden to measure the length of each signal. We emphasize that this difficulty—that the sender is

---

[1] Classical coding theory can be regarded as a special case when the signals $|u_i\rangle$'s are orthogonal.

[2] The most well-known quantum source coding scheme is a block coding scheme [1,2]. [The converse of this coding theorem was proven rigorously in [3].] To encode N signals *sequentially*, it requires $O(N^3)$ computational steps [4]. The encoding and decoding processes are far from instantaneous. Moreover, the lengths of all the codewords are the same.

ignorant of the length of the signals to be sent—is, in fact, very general. It appears in any distributed scheme of quantum computation. It is also highly analogous to the synchronization problem [5] in the execution of subroutines in a quantum computer: A quantum computer program runs various computational paths simultaneously. Different computational paths may take different numbers of computational steps. A quantum computer is, therefore, generally unsure whether a subroutine has been completed or not. We do not have a satisfactory resolution to those subtle issues in the general case. Of course, the sender can always avoid this problem by adding redundancies. However, such a prescription is highly inefficient and is self-defeating for our purpose of efficient quantum coding. For this reason, we reject such a prescription in our current discussion.

In the hope of saving resources, the natural next step to try is to stack the signals in line in a single-tape during the transmission. To greatly simplify our discussion we shall suppose that the read/write head of the machine is quantum mechanical with its location given by an internal state of the machine (this head location could be thought of as being specified on a separate tape). But then the second problem arises. Assuming a fixed speed of transmission, the receiver can never be sure when a particular signal, say the seventh signal, arrives. This is because the *total* length of the signals up to that point (from the first to seventh signals) is a quantum mechanical variable (i.e., it is in a superposition of many possible values). Therefore, Bob generally has a hard time in deciding when would be the correct instant to decode the seventh signal in an instantaneous quantum code.

Let us suppose that the above problem can be solved. For example, Bob may wait "long enough" before performing any measurements. We argue that there remains a third difficulty which is fatal for *instantaneous* quantum codes—that the head location of the encoder is *entangled* with the total length of the signals. If the decoder consumes the quantum signal (i.e., performs measurements on the signals) before the encoding is completed, the record of the total length of the signals in the encoder head will destroy quantum coherence. This decoherence effect is physically the same as a "which path" measurement that destroys the interference pattern in a double-slit experiment. One can also understand this effect simply by considering an example of N copies of a state $a|0\rangle + b|1\rangle$. It is easy to show that if the encoder couples an encoder head to the system and keeps a record of the total number of zeroes, the state of each signal will become impure. Consequently, the fidelity between the input and the output is rather poor.

III. STORAGE OF QUANTUM SIGNALS

Notice that the above problem is due to the requirement of instantaneousness. It would be helpful to have a general theorem to quantify the difficulties involved. Having failed that, we simply drop the requirement of instantaneousness and consider a simpler problem—the *storage* of quantum signals using a quantum analog of the Huffman coding algorithm. In this case, the decoding does not start until the whole encoding process is done and the code is not instantaneous in any sense. This immediately gets rid of the second (namely, when to decode) and third (namely, the record in the encoder head) problems mentioned in the last section. However, the first problem reappears in a new incarnation: The *total* length of say N signals is unknown and the encoder is not sure about the number of qubits that he should use.

A solution to this problem is to use essentially the law of large numbers. If N is large, then asymptotically the length variable of the N signals has a probability *amplitude* concentrated in the subspace of values between $N(\bar{L}-\delta)$ and $N(\bar{L}+\delta)$ for any $\delta > 0$ [1–3]. One can, therefore, truncate the signal tape into one with a *fixed* length say $N(\bar{L}+\delta)$. ['0's can be padded to the end of the tape to make up the number if necessary.] Of course, the whole tape is not of variable length anymore. Nonetheless, we will now demonstrate that this tape can be a useful component of a new coding scheme—which we shall call quantum Huffman coding—that shares many of the advantages of Huffman coding over block coding. In particular, assuming that quantum gates can be applied in *parallel*, the encoding and decoding of quantum Huffman coding can be done efficiently. While a sequential implementation of quantum source *block* coding [1–3] for N signals requires $O(N^3)$ computational steps [4], a parallel implementation of quantum Huffman coding takes only $O((\log N)^a)$ steps for some positive integer a.

We will now describe our new coding scheme, quantum Huffman coding, for the storage of quantum signals. As before, we consider a quantum source emitting a sequence of independent identically distributed quantum signals with a density matrix for each signal shown in Eq. (1) where p_i's are the eigenvalues. Considering Huffman coding for a classical source with probabilities p_i's allows one to construct a one-to-one correspondence between Huffman codewords h_i and the eigenstates $|\phi_i\rangle$. For parallel implementation, we find it useful to represent $|\phi_i\rangle$ by two pieces,[3] the first being the Huffman codeword, padded by the appropriate number of zeroes to make it into constant length,[4] $|0\cdots 0h_i\rangle$, the second being the length of the Huffman codeword, $|l_i\rangle$, where $l_i = \text{length}(h_i)$. We also pad zeroes to the second piece so that it becomes of fixed length $\lceil \log l_{\max} \rceil$ where $l_{\max}$ is the length of the longest Huffman codeword. Therefore, $|\phi_i\rangle$ is mapped into $|0\cdots 0h_i\rangle|l_i\rangle$. Notice that the length of the second tape is $\lceil \log l_{\max} \rceil$ which is generally small compared to n. The usage of the second tape is a small price to pay for parallel implementation.

A. Encoding

In this Section, we use the model of a quantum gate array for quantum computation. Now given $N = 2^r$ independent signals, the encoding can be done simply by divide-and-conquer. The first step is the merging of two signals into a single message. Let us introduce a message tape. For simplicity, we simply denote $|0\cdots 0h_{i_1}\rangle$ by $|h_1\rangle$, etc.

$$
\begin{aligned}
& |h_1\rangle|l_1\rangle|h_2\rangle|l_2\rangle|\mathbf{0}\rangle_{\text{tape}} \\
\xrightarrow{\text{swap}} \; & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle\ |0\cdots 0h_1\rangle_{\text{tape}} \\
\xrightarrow{\text{shift}} \; & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle\ |h_1 0\cdots 0\rangle_{\text{tape}}
\end{aligned}
$$

---

[3]The second piece contains no new information. However, it is useful for a massively parallel implementation of the shifting operations, which are an important component in our construction.

[4]The encoding process to be discussed below will allow us to reduce the total length needed for N signals.

$$\xrightarrow{\text{swap}} \ |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle \ |h_1 0\cdots 0 h_2\rangle_{\text{tape}}$$
$$\xrightarrow{\text{shift}} \ |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle \ |h_1 h_2 0\cdots 0\rangle_{\text{tape}} \ . \tag{2}$$

We remark that the swap operation between any two qubits can be done efficiently by using an array of three XOR's with the two qubits alternately used as the control and the target. What is not entirely obvious to show is that the quantum shift operations in steps 2 and 4 can be efficiently implemented *in parallel* in a quantum gate array model. [We will discuss this point in the Appendix.] Now the encoder keeps the original length tape for *each* signal as well as the message tape for two messages, i.e., $|l_1\rangle|l_2\rangle|h_1 h_2 0\cdots 0\rangle_{\text{tape}}$. Notice that it is relatively fast to compute the length $l_1 + l_2$ of the two messages from l_1 and l_2. This completes our discussion on how to merge two messages into one in polynomial time.

Without much loss of generality, we suppose that the total number of messages is $N = 2^r$ for some positive integer r. We propose to encode by divide and conquer. Firstly, we divide the messages into pairs and apply the above merging procedure to each pair. The merging effectively reduces the total number of messages to 2^{r-1}. We can repeat the above process. Therefore, after r applications of the merging procedure, we obtain a single tape containing all the messages (in addition to the various length tapes containing the length information).

More concretely, at the end the encoder obtains $|l_1\rangle|l_2\rangle\cdots|l_N\rangle|h_1 h_2\cdots h_N 0\cdots 0\rangle_{\text{tape}}$ in only $O((\log N)^a)$ computational steps for some positive integer a. Finally, the encoder truncates the message tape: He keeps only say the first $N(\bar{L}+\delta)$ qubits in the message tape $|h_1 h_2\cdots h_N 0\cdots 0\rangle_{\text{tape}}$ for some $\delta > 0$ and throws away the other qubits. This truncation minimizes the number of qubits needed. The only overhead cost compared to the classical case is the storage of the length tapes of the individual signals. This takes only $N\lceil\log l_{\max}\rceil$ qubits.[5]

B. Decoding

Decoding can be done by adding an appropriate number of qubits in the zero state $|0\rangle$ behind the truncated message tape and simply running the encoding process backwards (again with only $O((\log N)^a)$ computational steps).

What about fidelity? The key observation is the following: Just like the case of Schumacher's noiseless quantum coding theorem [1–3], the "typical" space is contained in the tensor product space of the stored qubits, (i.e., the first $N(\bar{L}+\delta)$ qubits) with a *fixed* state $|0\cdots 0\rangle$ for the remaining qubits. Therefore, the truncation and subsequent replacement of the discarded portion by $|0\cdots 0\rangle$ still lead to a high fidelity in the decoding.

In conclusion, we have constructed an explicit parallel encoding and decoding scheme for the storage of N independent and identically distributed quantum signals that asymptotically uses only $O((\log N)^a)$ computational steps and $N(\bar{L}+\delta+\lceil\log l_{\max}\rceil)$ qubits for storage where $\bar{L}$ is the average length of the Huffman coding for the classical coding problem for the

---

[5]Further optimization may be possible. For instance, if $\log l_{\max}$ is large, one can save storage space by repeating the procedure, i.e., one can now use quantum Huffman coding for the problem of storing the quantum signals $|l_i\rangle$s.

set of probabilities given by the eigenvalues of the density matrix of each signal. Here δ can be any positive number and $l_{\max}$ is the length of the longest Huffman codeword.

IV. COMMUNICATION

We now attempt to use the quantum Huffman coding for communication rather than for storage of quantum signals. By communication, we assume that Alice receives the signals *one by one* from a source and is compelled to encode them one by one. As we will show below, the number of qubits required is slightly more, namely $N(\bar{L} + \delta + 2\lceil \log l_{\max} \rceil)$. The code that we will construct is not instantaneous, but Alice and Bob can pay a small penalty in stopping the transmission any time.

A. Encoding

The encoding algorithm is similar to that of Section 3 except that the signals are encoded one by one. More concretely, it is done through alternating applications of the swap and shift operations.

$$
\begin{aligned}
& |h_1\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle|\mathbf{0}\rangle_{\text{tape}} \\
\xrightarrow{\text{swap}}\ & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle\ |0\cdots 0h_1\rangle_{\text{tape}} \\
\xrightarrow{\text{shift}}\ & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle\ |h_1 0\cdots 0\rangle_{\text{tape}} \\
\xrightarrow{\text{swap}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle\ |h_1 0\cdots 0h_2\rangle_{\text{tape}} \\
\xrightarrow{\text{shift}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle\ |h_1 h_2 0\cdots 0\rangle_{\text{tape}} \\
& \cdots \\
\xrightarrow{\text{shift}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|\mathbf{0}\rangle|l_N\rangle\ \ |h_1 h_2\cdots h_N 0\cdots 0\rangle_{\text{tape}}\ . \qquad (3)
\end{aligned}
$$

Even though the encoding of signals themselves are done one-by-one, the shifting operation can be speeded up by parallel computation. Indeed, as shown in the Appendix, a parallel implementation can achieve the shifting operation in $O((\log N)^a)$ steps. If a sequential implementation is used instead, the shifting operation still requires only $O(N(\log N)^a)$ steps for the following reason. At each time step of a parallel implementation, only $O(N)$ steps are implemented. Since there are only $O((\log N)^a)$ steps in a parallel implementation of the shifting operation, multiplying the two gives $O(N(\log N)^a)$ steps for a sequential implementation of a shifting operation.

Now the encoding of the N signals in quantum communication is done sequentially, implying $O(N)$ applications of the shifting operation. Therefore, with a parallel implementation of the shifting operation, the whole process takes $O(N(\log N)^a)$ steps. With a sequential implementation, it takes $O(N^2(\log N)^a)$ steps.

B. Transmission

Notice that the message is written on the message tape from left to right. Moreover, starting from left to right, the state of each qubit once written remains unchanged throughout

set of probabilities given by the eigenvalues of the density matrix of each signal. Here δ can be any positive number and $l_{\max}$ is the length of the longest Huffman codeword.

IV. COMMUNICATION

We now attempt to use the quantum Huffman coding for communication rather than for storage of quantum signals. By communication, we assume that Alice receives the signals *one by one* from a source and is compelled to encode them one by one. As we will show below, the number of qubits required is slightly more, namely $N(\bar{L} + \delta + 2\lceil \log l_{\max} \rceil)$. The code that we will construct is not instantaneous, but Alice and Bob can pay a small penalty in stopping the transmission any time.

A. Encoding

The encoding algorithm is similar to that of Section 3 except that the signals are encoded one by one. More concretely, it is done through alternating applications of the swap and shift operations.

$$
\begin{aligned}
& |h_1\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle|\mathbf{0}\rangle_{\text{tape}} \\
\xrightarrow{\text{swap}}\ & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle\ |0\cdots 0h_1\rangle_{\text{tape}} \\
\xrightarrow{\text{shift}}\ & |\mathbf{0}\rangle|l_1\rangle|h_2\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle\ |h_1 0\cdots 0\rangle_{\text{tape}} \\
\xrightarrow{\text{swap}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle\ |h_1 0\cdots 0h_2\rangle_{\text{tape}} \\
\xrightarrow{\text{shift}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|h_N\rangle|l_N\rangle\ |h_1 h_2 0\cdots 0\rangle_{\text{tape}} \\
& \cdots \\
\xrightarrow{\text{shift}}\ & |\mathbf{0}\rangle|l_1\rangle|\mathbf{0}\rangle|l_2\rangle\cdots|\mathbf{0}\rangle|l_N\rangle\ \ |h_1 h_2\cdots h_N 0\cdots 0\rangle_{\text{tape}}\ . \qquad (3)
\end{aligned}
$$

Even though the encoding of signals themselves are done one-by-one, the shifting operation can be speeded up by parallel computation. Indeed, as shown in the Appendix, a parallel implementation can achieve the shifting operation in $O((\log N)^a)$ steps. If a sequential implementation is used instead, the shifting operation still requires only $O(N(\log N)^a)$ steps for the following reason. At each time step of a parallel implementation, only $O(N)$ steps are implemented. Since there are only $O((\log N)^a)$ steps in a parallel implementation of the shifting operation, multiplying the two gives $O(N(\log N)^a)$ steps for a sequential implementation of a shifting operation.

Now the encoding of the N signals in quantum communication is done sequentially, implying $O(N)$ applications of the shifting operation. Therefore, with a parallel implementation of the shifting operation, the whole process takes $O(N(\log N)^a)$ steps. With a sequential implementation, it takes $O(N^2(\log N)^a)$ steps.

B. Transmission

Notice that the message is written on the message tape from left to right. Moreover, starting from left to right, the state of each qubit once written remains unchanged throughout

the encoding process. This decoupling effect suggests that rather than waiting for the completion of the whole encoding process, the sender, Alice, can start the transmission immediately after the encoding. For instance, after encoding the first r signals, Alice is absolutely sure that at least the first $rL_{\min}$ (where $L_{\min}$ is the minimal length of each codeword) qubits on the tape have already been written. She is free to send those qubits to Bob immediately. There is no penalty for such a transmission because it is easy to see that the remaining encoding process requires no help from Bob at all. [Note that in the asymptotic limit of large r, after encoding r signals, Alice can even send $r(\bar{L}-\epsilon)$ qubits for any $\epsilon > 0$ to Bob without worrying about fidelity.]

In this scheme for communication, we propose that Alice makes a duplicate of the length tape for each signal, i.e., she evolves

$$\begin{aligned} & |l_i\rangle \otimes |0\rangle \\ \rightarrow\ & |l_i\rangle \otimes |l_i\rangle \end{aligned} \tag{4}$$

and sends a copy to Bob immediately afterwards. Therefore, an additional number of $N\lceil \log l_{\max} \rceil$ qubits are needed for communication in comparison to storage.

C. Decoding

With the length information of each signal and the received qubits, Bob can start the decoding process before the whole transmission is complete *provided that* he does not perform any measurement at this moment. For instance, having received $rL_{\min}$ qubits in the message tape from Alice, Bob is almost sure that at least $s = \lfloor rL_{\min}/L_{\max} \rfloor$ signals have already arrived. He can decode those signals immediately using the length information of each signal. This decoding process is rather straightforward and we will skip its description here.

D. Measurements

The important observation is, however, the following: If Bob were to perform a measurement on his signals now, he would find that his signals are of poor fidelity. The reason behind this has already been noted in Section 2. Even though the subsequent encoding process does not involve Bob's system, there is still entanglement between Alice and Bob's systems. More specifically, the shifting operations in the remaining encoding process by Alice require explicitly the information on the total length of decoded signals. Before Bob performs any measurement on his decoded signals, it is, therefore, crucial for Alice to disentangle her system first.

Suppose in the middle of their communication in which Bob has already received $K\bar{L}$ qubits from Alice, Bob suddenly would like to perform a measurement on his signals. He shall first inform Alice of his intention. Afterwards, one way to proceed is the following: They choose some convenient point, say the m-th signal, to stop and consider quantum Huffman coding for only the first m signals and complete the encoding and decoding processes.

We shall consider two subcases. In the first subcase, the number m is chosen such that the m-th signal is most likely still in the sender (Alice)'s hands. [e.g. $m > K + O(\sqrt{K})$ in the asymptotic limit.] The sender Alice now disentangles the remaining signal from the first

m quantum signals by applying a quantum shifting operation. She can now complete the encoding process for quantum Huffman coding of the m signals and send any un-transmitted qubits to Bob. In the asymptotic limit of large K, $O(\sqrt{m})$ qubits of forward transmission (from Alice to Bob) are needed. (The number of computational steps needed is polynomial in $\log m$ *if* a parallel implementation of a quantum gate array is used.)

In the second subcase, the number m is chosen such that the m-th signal is most likely already in the receiver (Bob)'s hands. [e.g. $m < K - O(\sqrt{K})$ in the asymptotic limit.] The receiver Bob now attempts to disentangle the remaining signals from the first m quantum signals by applying a quantum shifting operation. Of course, he needs to shift some of his qubits back to Alice. This asymptotically amounts to $O(\sqrt{m})$ qubits of *backward* communication. This is a penalty one must pay for this method. We remark that the shifting operation can be done rather easily in distributed quantum computation between Alice and Bob. This is a non-trivial observation because the number of qubits to be shifted from Alice to Bob is itself a quantum mechanical variable. This, however, does not create much problem. Bob can always communicate with Alice using a bus of fixed length. For example, he applies local operations to swap the desired quantum superposition of various numbers of qubits from his tape to the bus, sends such a bus to Alice, etc.

In the above discussion, we have focused on the simple case when Bob would like to perform a measurement on the whole set of the first m signals. Suppose Bob is interested only in a particular signal say the m-th one, but not the others. There exists a more efficient scheme for doing it. We shall skip the discussion here.

V. CONCLUDING REMARKS

We have successfully constructed a Huffman-coding inspired scheme for the storage of quantum information. Our scheme is highly efficient. The encoding and decoding processes of N quantum signals can be done *in parallel* in computational steps polynomial in $\log N$. (If parallel machines are unavailable, as shown in subsection IV A our encoding scheme will still take only $O(N^2(\log N)^a)$ computational steps for a sequential implementation. In contrast, a naive implementation of Schumacher's scheme will require $O(N^3)$ computational steps.) This massive parallelism is possible because we explicitly use another tape to store the length information of the individual signals. The storage space needed is asymptotically $N(\bar{L} + \delta + \lceil \log l_{\max} \rceil)$ where $\bar{L}$ is the average length of the corresponding classical Huffman coding problem for the density matrix in the diagonal form, δ is an arbitrary small positive number and $l_{\max}$ is the length of the longest Huffman codeword.

We also considered the problem of using quantum Huffman coding for communication in which case Alice encodes the signals one by one. $N(\bar{L} + \delta + 2\lceil \log l_{\max} \rceil)$ qubits are needed. With a parallel implementation of the shifting operation, $O(N(\log N)^a)$ computational steps are needed. On the other hand, with a sequential implementation, $O(N^2(\log N)^a)$ computational steps are needed. In either case, the code is not instantaneous, but, by paying a small penalty in terms of communication and computational costs, Alice and Bob has the option of stopping the transmission and Bob may then start measuring his signals.

More specifically, while the receiver Bob is free to decode the signals, he is not allowed to measure them until the sender Alice has completed the encoding process. This is because Alice's encoder head generally contains the information of the total length of the signals. In

other words, its state is entangled with Bob's signals. Therefore, whenever Bob would like to perform a measurement, he should first inform Alice and the two should proceed with disentanglement. We present two alternative methods of achieving such disentanglement at small penalties of communication and computational costs.

Since real communication channels are always noisy, in actual implementation source coding is always followed by encoding into an error correcting code. Following the pioneering work by Shor [7] and independently by Steane [8], various quantum error correcting codes have been constructed. We remark that quantum Huffman coding algorithm (even the version for communication) can be immediately combined with the encoding process of a quantum error correcting code for efficient communication through a noisy channel.

As quantum information is fragile against noises in the environment, it may be useful to work out a fault-tolerant procedure for quantum source coding. The generalizations of other classical coding schemes to the quantum case are also interesting. According to M. A. Nielsen, there exist universal quantum data compression schemes motivated by the well-known Lempel-Ziv compression algorithm for classical information [9].

Notes Added: After the completion of this work, we were informed by A. M. Steane that our work overlaps with some unpublished work by B. Schumacher [10]. We thank A. M. Steane for bringing this to our attention and B. Schumacher for helpful conversations.

VI. ACKNOWLEDGMENT

One of us (H.-K. Lo) thanks D. P. DiVincenzo, J. Preskill and T. Spiller for helpful discussions. This work is supported in part by EPSRC grants GR/L91344 and GR/L80676, by a Lee A. DuBridge Fellowship and by DARPA under Grant No. DAAH04-96-1-0386 through the Quantum Information and Computing (QUIC) Institute administered by ARO and by U.S. Department of Energy under Grant No. DE-FG03-92-ER40701.

APPENDIX: SHIFTING

The efficiency of the encoding and decoding processes of quantum Huffman coding presented in this paper relies heavily on an efficient algorithm for quantum shifting. Here we will first demonstrate an efficient (i.e., polynomial in $\log N$) classical algorithm for shifting $N = 2^k$ objects in a circle by one position. Afterwards, we generalize it to a shift by an arbitrary number of positions. Finally, we discuss its generalization to the quantum case.

First, consider a shift of 2^k objects by one position. Our algorithm is best understood by an example. Suppose eight objects (i.e., $k = 3$) labelled by $1, 2, \cdots, 8$ arranged in a line. The first step is a parallel implementation of the interchange operations of neighbouring objects (1 2), (3 4), (5 6) and (7 8). This maps the line $12345678 \rightarrow 21436587$. The second step is a parallel implementation of (1 3) and (5 7). This maps $21436587 \rightarrow 23416785$. The final step is (1 5), which maps $23416785 \rightarrow 23456781$. Therefore, in $\log 8 = 3$ steps we have succeeded in shifting 8 objects by one position. Now, it is simple to construct an algorithm for shifting 2^k objects by 2^i positions using only $k - i$ steps: One just divides the objects into 2^i subsets according to their values modulo 2^i and applies the shifting by one step on each of the individual 2^i subsets containing 2^{k-i} objects.

Now, for a general shift by r positions, one simply expands r in binary. A sequence of shifting operations by 2^i is now applied for those non-zero entries in the binary expansion. Hence, at most $(\log N)^2$ parallel computational steps are needed.

One can generalize the classical shifting algorithm to a quantum one by adding a control to each shifting (by 2^i) operation. We remark that, since the length tapes $|l_i\rangle$ contain all the information on the amount of shifting needed, a parallel implementation is feasible. This is the key reason why we insist on retaining the length tapes in the first place.

REFERENCES

[1] B. Schumacher, Phys. Rev. **A51**, 2738 (1995).
[2] R. Jozsa and B. Schumacher, J. Mod. Opt. **41**, 2343 (1995).
[3] H. Barnum, C. A. Fuchs, R. Jozsa, and B. Schumacher, Phys. Rev. **A 54**, 4707 (1996).
[4] R. Cleve and D. P. DiVincenzo, Phys. Rev. **A54**, 2636 (1996).
[5] Y. Shi, http://xxx.lanl.gov/abs/quant-ph/9705017.
[6] R. Jozsa, Journal of Modern Optics **41**, 2315 (1994).
[7] P. W. Shor, Phys. Rev. **A52**, R2493 (1995).
[8] A. M. Steane, Phys. Rev. Lett. **77**, 793 (1996).
[9] M. A. Nielsen, private communications.
[10] B. Schumacher, private communications on unpublished work.