Abstract— The conventional encryption methods are not appropriate for image cryptography as they cover huge volume of data and the pixels have a strong correlation. Currently, data encryption and encoding have become more critical and challenging, in multimedia communication. The rapid developments in digital communication have made wide use of multimedia data. The basic characteristics of multimedia communication system like security, high transmission rate, low bandwidth, redundancy of data and low storage capacity makes basic encryption and compression algorithms mandatory. This paper discusses the basic Baptista chaotic system and its variants made by incorporating Huffman encoding. This image encryption algorithm optimizes the compression performance in Baptista’s chaotic system. Result analysis proves that the compression performance and encryption speed of the modified system is much better. This paper made a proposal for implementing this optimization algorithm in cloud computing. So that any organization can exploit the cloud computing as an optimization framework for the large amount of data sharing by preserving better security and compression performance.

Keywords— Encryption, Cloud computing, Compression, Cryptography, Chaos, Logistic map.

I. INTRODUCTION

Cloud computing is a group of services including the hardware and operating system infrastructure and virtualization components. Cloud architectures are established in an on-demand manner. That is, the resources are dynamically assigned to a user according to his request, and surrendered after the job is done. Cloud computing enables us to access shared resources and manipulation of large amounts of stored data over public or private networks. By the optimization features like resource and data sharing, it became very popular in the internet community. It helps to reduce the cost of IT infrastructure and maintenance by meeting dynamic unpredictable demand on resources. In cloud computing, information is distributed over the network. So the security of these large amounts of data become a main challenge in this field. To secure the data, it has to apply some cryptographic algorithms.

In the present era of multimedia communication networks, cryptography has achieved special consideration. It is the science of providing the security features of information in any communication system. Traditional cryptographic techniques are based on the number theory. Pecorra and Carrol discovered chaotic synchronization principles. The chaotic system has the manners of a nonlinear dynamic system, which seemingly appearances as random [4]. But it is resulting from the defining deterministic processes for particular system parameters. The basic properties of chaotic schemes [11] are sensitivity on initial conditions and control parameters, which provide secure communication methods. Nowadays chaos based cryptographic systems seems like a substitute to the existing techniques such as the Rivest-Shamir-Adleman algorithm (RSA), Elliptic Curve Cryptography etc. The chaotic cryptographic scheme can be implemented in the simplest form of one or two dimensional systems represented by discrete chaotic maps. This chaotic method can generate complicated dynamics with the use of chaotic digital maps. This is desirable in cryptographic algorithms.

There exists a strong linking between cryptography and chaos. Unpredictable chaotic orbits are generated with the deterministic equations of chaotic systems, as they are extremely sensitive to its initial conditions. The chaotic properties like sensitivity to initial values, ergodicity, mixing, etc., are comparable with the confusion-diffusion properties of conventional cryptography.

By the sensitivity nature of chaos to initial conditions, initially very near trajectories may diverge exponentially in a small period of time. Without the initial information about the system, long term forecast of the upcoming statuses of the chaotic system is impossible. By the ergodicity property, a path in phase space comes arbitrarily near to its former status. This reflects that the system is eventually limited to its spatial object made up of an attractor. The density of this attractor is time invariant. By mixing property, a small interval of the system is much better. This paper made a proposal for implementing this optimization algorithm in cloud computing.
In recent years, compressing the files before transmitting has gained a lot of interest with a fast growth of multimedia data through an open network. The method of decreasing the size of a data file is often denoted as data compression. Compression techniques are used to decrease the quantity of data for saving the storage space volume. It is called source coding as the data source is encoded before transmission. Compression is advantageous because it reduces resources essential to store and transmit data.

In data compression, information is represented with less bits than the original. Compression is of two types, lossless or lossy. Lossless means no data is missing in the decompression, while it is kept in the compressed format. Lossy means some data is lost at the time of decompression. Lossless data compression is classified into Run-Length, Huffman and Transformation. Transformation is again divided into Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT). Lossy data compression is classified into Joint Photographic Experts Group (JPEG), Moving Picture Experts Group (MPEG) and MPEG Audio Layer 3. Huffman coding is a variable distance coding. Since it provides increased compression rate, it is commonly used as a compression technique during transmission of data.

The rest of this paper is organized into 4 sections. The section 2 will explain the existing methods of cloud computing, encryption and compression method. The section 3 deals with the integration of compression methods in the search based chaotic encryption scheme. This explains in detail about the new variant of Baptista method proposed by Wong et al. The section 5 deals with the result analysis.

II. EXISTING METHOD

A. Cloud Computing

Cloud computing is a technology which enables universal access to shared data and services [1] through the internet. Cloud computing became very popular with the large network facility and other service oriented environment. Any organization can take benefit from this optimizing technology. By optimizing IT infrastructure, it helps the company to cut costs of the organization. Virtualization technology enables cloud computing to increase the speed of computing through the large infrastructure utilization. Cloud computing is a type of grid computing. Other than the conventional parallel computing, it provides a dynamic platform for the parallel applications with reasonable price.

However, security aspects of the data in the cloud computing environment need much more concern [3]. Since large amounts of data is shared through an open network. Lack of protection of the data is the major challenge in the cloud computing [13]. The complexity of dealing with the information security makes uncertainty in this field. The best solution for this problem is the introduction of encryption algorithms in this area. Therefore, this paper proposed to implement the encryption algorithm in a cloud computing system. The encryption can prevent unauthorized data accessing in the open network. This is one of the best and most popular solutions for securing confidential information. Other than the conventional encryption method, the search based chaotic encryption algorithm can provide better security features.

B. Encryption

In 1998, M.S. Baptista had proposed a chaos-based crypto system which became very popular in the area of chaotic cryptography. This method, inputs a text data composed by some alphabet as the plaintext. Taking the benefit of ergodicity property, the ciphertext is created from the plaintext. The system is limited to a group of points called an attractor. The whole of the attractor is divided into S sites. In this method, S = 256. The size of each site can be calculated using equation (1).

\[ e = \frac{x_{\text{max}} - x_{\text{min}}}{S} \]  

(1)

Where \( x_{\text{max}} \) and \( x_{\text{min}} \) are the corresponding upper and lower value of the attractor. The entire phase space of the given map is distributed into a number of equal size partitions. Each partition, \( e \) interval, map to a plaintext symbol. The partitioned phase space along with the corresponding plaintext mapping is considered as a lookup table for encryption. A pseudo random sequence is generated by a chaotic map is used for searching in the lookup table. Baptista method uses logistic map as the chaotic map. This is a one-dimensional map proposed by R.M. May as in equation (2).

\[ x_{n+1} = bx_n(1 - x_n) \]  

(2)

Where \( x_n \in [0,1] \) is the output at discrete time \( n=0, 1, 2, \ldots \) and control parameter \( b \) should be a real number between 3.6 and 4. The above equation provides its trajectory, leaving from an initial condition \( x_0 \) reach an \( e \)-interval associated with that character. Fig. 1 shows a schematic representation of the way of associating the \( S \) units alphabet with the \( S^e \) intervals.

A secret chaotic trajectory is generated from the parameters and initial condition. This will search the corresponding part of the plaintext symbol which has to be encrypted. The number of iterations of the logistic map will be the length of the searching trajectory. This number of iterations will be the ciphertext.

In the decryption process, the similarly secret chaotic search path is redeveloped. It can recover the plaintext symbols with the identical secret key and the lookup table. The Baptista-type search based chaotic crypto system has low compression performance. The ciphertext is generally around 1.5 to 2 times of the plaintext length. So it needs to implement a new method which has better compression performance than the Baptista method. In the following section, a Huffman
encoding scheme is being discussed. Reference [5] used this for incorporating compression in Baptista method.

C. Compression

The In 1952, David Huffman developed a lossless Huffman coding system for data compression. Based on the possibility of occurrence of every symbol, it will generate a variable length code. It reduces the size of data by giving a short code for more possible symbols and longest code for less probable symbols. In the 1940s, Claude Shannon and Robert Fano proposed independently a coding method known as Shannon-Fano method. They constructed codes from the topmost to bottom, but in Huffman coding it is from the bottom to top.

Huffman coding technique consists of simple steps of sorting and merging of symbols. It constructs a binary tree starts from the bottom. At first it wants to sort all symbols based on the chances of occurrence in descending order. In the beginning all the symbols are considered as the leaf nodes. A binary tree is built through several steps. In each step, two smallest possible symbols take into consideration and combine them into a single one. Add probabilities of each symbol and it will be the possibility of the combined one. Continuing like this, at the end there will be only one symbol. All the symbols are merged to become one single symbol. To calculate the code words of each symbol, it wants to traverse the binary tree from root to a leaf node. It requires to assign arbitrarily a 0 to the left edge and 1 to the right edge.

<table>
<thead>
<tr>
<th>SI No</th>
<th>Code words and Symbols</th>
<th>Symbols</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>0.125</td>
<td></td>
</tr>
</tbody>
</table>

The process can be explained with an example as in Fig. 2. As in the Table I, four symbols are given with their equivalent probabilities. At first, sort the symbols with their probability of occurrences in descending order. Here C and D have smallest probability of 0.125. So it needs to take that two symbols as leaf nodes and add its probability. So 0.25 will be the frequency of the new merged symbol CD. The symbol B has the probability 0.35. So merge B with CD to get new merged symbol BCD with probability 0.60. This BCD can combine with the symbol A to get ABCD as the final root node with a probability 1.0. Thus the binary tree will be completed at the bottom to up manner. By assigning 0 bits at the left edge and 1 bit in the right edge, the code words of the four symbols can be determined as in the Table II. Average size of the code word, i.e. Bit rate (R) can be calculated by (3).

\[ R = \frac{\sum_{i=1}^{n} P_i l_i}{(3)} \]

Where \( P_i \) is the probability of each symbol and \( l_i \) is the length of the code word of each symbol.

\[ R = 4*1+.35*2+.125*3+.125*3=1.85 \text{ bits/symbol.} \]

The basic Huffman coding method is static and it will take too much time for compression. To generate the code word of all the symbols, it may take two passes. One is for counting the frequencies of entire symbols and other is for the compression. To solve this problem, the dynamic Huffman coding is developed by Faller and Gallager with substantial improvements by Knuth.
In this, the Huffman tree will be empty at the initial state. It will modify the tree in each step while reading the symbols. The first symbol in the input is added to the binary tree and will allot a code for it. At the time of the next visit, the current code is written to the output and the frequency will be incremented by one. This will modify the tree and check whether it is still a Huffman tree. Otherwise, it requires rearranging. To modify a tree, check each time, while inputting a symbol. Gallager introduced a theorem by which a full tree is said to be Huffman, if it satisfies the sibling property. The frequency of symbols in each level from left to right and from bottom to top, must be sorted in non-descending order. Thus the root has the highest frequency and bottom left has the lowest probability.

The basic algorithm consists of comparison and swapping of symbols. At first it compares the probability of the present node to the succeeding nodes. If the probability of the present node is lower than its successor, there won’t be any change. Otherwise, it requires to swap with the one which has the smallest frequency. If there exists more than one node with small identical frequency than the present node, swap it with the last node in the group of the same frequency. The current node should not be swapped with its parent node. Then increment the occurrence of the current node by one. Correspondingly increment the frequencies of its parents. Until the current node becomes the root node, the process has to be repeated.

1) Adaptive Huffman Coding

The adaptive Huffman coding [10] can be explained in detail with an example as in Fig. 3. The tree contains four symbols, A, B, C and D and its frequencies. 7 symbols in the tree are already processed. Here symbol A is taken as the present node and the modification of tree considers the updating status of only symbol A. At each step, checking of sibling property of binary tree has to be conducted. If it satisfies the property, it can say that it is a Huffman tree.

Left binary tree in Fig. 3 shows the present frequency level of 4 symbols and right binary tree shows the final binary tree. In the first step, frequency of symbol A is incremented by 1 and it becomes 2. Now it satisfies the sibling property. So it does not want swapping. The only change in the tree is the corresponding increment in all the parent nodes of A. In the next step, frequency of A becomes 3. Frequency of A exceeds its successor B, C and D. Symbols B, C and D have the same frequency. So it needs to swap with the last one of the group of similar frequency. So this needs a swapping of D and A. Otherwise, it violates the sibling property.

In the next step, frequency of A is incremented from 3 to 4. Since all nodes satisfies the sibling property, swapping is not required. Now the frequency of A is 4 and which is equal to that of its successor in the immediate top level. So they should be swapped. In the succeeding step, frequency of A is incremented from 4 to 5. This will be the final tree, for the updating of frequency level of symbol A.

The adaptive Huffman tree can be implemented in the process of reading the file itself. It offers the facility to update the tree dynamically. So there is no time wastage in the process of compression. While in static model updating, it requires two passes – one for reading the entire input and one for creating the binary tree. So in practical applications, it is better to use the adaptive Huffman encoding rather than the static one. There are so many variations proposed by recent researchers for the adaptive Huffman encoding scheme. The idea behind all of them are same. Wong et al incorporated this compression technique in the Baptista method to improve the compression performance.

III. PROPOSED METHOD

Reference [7] proposed some variation of Baptista type search based chaotic crypto systems. Baptista used a static lookup table for the encryption, while a dynamic lookup table [6] based on the possibility of occurrence of the symbols are used in the modified scheme. As in the basic Baptista system, it maps the plaintext in a lookup table. The difference is that the former is static and the latter is dynamic. At first, scan the whole plaintext symbols and find out the frequency of every symbol. The whole phase space is divided and based on the chances of occurrence of the symbol, each one is assigned into the partitions. It can be better explained with an example as follows. Take 5 symbols A, B, C, D and E. Each one will be assigned the partition as per the frequency in the table. Total 256 partitions are there on a lookup table.

<table>
<thead>
<tr>
<th>SI No</th>
<th>Partitions based on probability of Symbols</th>
<th>Probability</th>
<th>No of Partitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>0.5</td>
<td>128</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>0.25</td>
<td>64</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>0.125</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>0.0625</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>0.0625</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>256</td>
</tr>
</tbody>
</table>

As given in Table III, based on the possibility of occurrence of the symbols, here 128 partitions will map to A, 64 will map to B, 32 will map to C and 16 each will map to both the symbol D and E. The corresponding dynamic lookup table is shown by Fig. 4. In the following step, it has to be checked for the plaintext symbol which is in the corresponding lookup table or not. If it is there, iterate the chaotic logistic map up to which it will reach the relevant parts of the dynamic lookup table. Here it also uses the logistic map as the underlying chaotic map to produce a
A pseudorandom sequence as in Baptista method, \( x_n \in [0, 1] \) is the output at discrete time \( n = 0, 1, 2 \) etc. and the parameter \( b \) should be a real number between 3.6 and 4. The control parameter, \( b \) and the initial condition, \( x_0 \) are taken as the secret key in this encryption method. Since it is a symmetric key encryption scheme, it uses the same key for both the process of encryption and decryption.

The intermediate sequence should be masked by the binary mask sequence \( m \) obtained at the time of iteration before transmission. In this process, both the binary mask sequence and the intermediary sequence are separated into 32-bit blocks. The ciphertext of the block, ‘i’ is given by equation (4)

\[
c = (r_{i+1} + m_{c_{i-1}} + r_{i-1,mod(L/4)} + c_{i-1})mod2^{32}
\]

Where \( c_i \) and \( r_i \) are the \( i^{th} \) 32-bit block in the ciphertext and intermediate sequences, respectively. The location of the mask block for \( r_i \) is determined by \( c_{i-1,mod(L/4)} \), where \( L \) is the plaintext length in bytes. To encrypt the first block \( c_0 \), \( c_0-1 \) is required, which is calculated by the secret key. If this covering process is performed only once, a tiny change in the last bit of the plaintext sequence may affect only a minor portion at the end of the ciphertext sequence. Therefore, it should be performed at least twice consecutively so as to ensure that a small change in any location of the plaintext sequence spreads over the whole ciphertext sequence. The extra block \( r_{i+1} \) required at the end, is set to \( c_{0-1} \).

The advantage of this scheme compared with the Baptista method is that it can represent ciphertext with a small number [8]. Because the number of iterations needed for more probable symbols will be less. The chance of getting searched will be high for those symbols which has a number of partitions. In this, the more probable symbols can only be encrypted. The fewer probable symbols will be encrypted by mask mode by a pseudo random bit stream generated from the logistic map.

IV. Result

A. Compression Ratio

The standard files from the Calgary Corpus are used for calculating the compression performance. There are 7 separate files of different types. Two simulation configurations are chosen. The first column, only 16 higher probable plaintext symbols are selected and they are all mapped to one table. In the second column, 128 higher probable plaintext symbols are selected. Then they are mapped to 16 tables of 8 symbols per each. The compression ratio (CR) calculated by equation (5) is listed in Table 4 for different test files.

\[
CR = (\text{ciphertext} / \text{plaintext}) \times 100 \%
\]

Table IV shows that all files can be compressed using the two configurations. However, the compression performance of the second configuration (16 maps, eight symbols, each) is better for most of the files.

**Table 4**: Compression Ratio

<table>
<thead>
<tr>
<th>File Type</th>
<th>Compression Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calgary corpus 1</td>
<td>80%</td>
</tr>
<tr>
<td>Calgary corpus 2</td>
<td>90%</td>
</tr>
<tr>
<td>Calgary corpus 3</td>
<td>95%</td>
</tr>
<tr>
<td>Calgary corpus 4</td>
<td>98%</td>
</tr>
<tr>
<td>Calgary corpus 5</td>
<td>99%</td>
</tr>
<tr>
<td>Calgary corpus 6</td>
<td>99.5%</td>
</tr>
<tr>
<td>Calgary corpus 7</td>
<td>100%</td>
</tr>
</tbody>
</table>
B. Plaintext Sensitivity

To test the plaintext sensitivity, a bit is changed at different locations of the plaintext sequence, and then encoded by the same key under the all-mask mode. The two resultant ciphertext sequences are compared bit-by-bit, and the ratio of bit change is calculated. As in Table V, bit change at the beginning of the plaintext calculated as 50.00%, bit change in the middle of the plaintext as 50.04% and at the end of the plaintext as 50.01%. This result shows that the ciphertext is very sensitive to the plaintext even for a bit change at different positions. Two rounds of masking are done and so change spreads over the entire ciphertext sequence. The security of this scheme is high for both the all-mask mode and hybrid mode. For hybrid mode, the ciphertext block length is not fixed due to the occasional insertion of the variable-length Huffman code. It is difficult to identify correctly the bits corresponding to a Huffman code. Moreover, decoding a bit stream of Huffman codes is found difficult without any knowledge about the Huffman coding table. Therefore, the hybrid mode is considered more secure than the all-mask mode.

### TABLE V. PLAINTEXT SENSITIVITY

<table>
<thead>
<tr>
<th>Plaintext Sensitivity Analysis</th>
<th>Ratio of Bit Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning</td>
<td>50.00%</td>
</tr>
<tr>
<td>Middle</td>
<td>50.04%</td>
</tr>
<tr>
<td>End</td>
<td>50.01%</td>
</tr>
</tbody>
</table>

V. CONCLUSION

Existing approaches for simultaneous chaotic encryption and compression schemes are encryption oriented. Compression is not the main concern, and so, some of them suffer from the low compression performance. On the contrary, algorithm implemented by K.W Wong builds a chaotic cryptosystem that was designed for both encryption and compression. Results indicates that all the standard test files can be compressed to a satisfactory degree, and the ciphertext is very sensitive to a tiny change in the key or the plaintext. Therefore, the compression capability is achieved while the security is maintained. This system also ensures that the ciphertext is not lengthier than the plaintext. Thus it has attained the basic optimization criteria for an encryption algorithm and can be applied in cloud computing. Cloud computing enables an organization to work with minimum infrastructure of resource and data. By giving an online facility to manage its runtime demand of resources and data, cloud computing became popular in its field. By giving a better security feature along with a better compression performance, IT team can exploit it in a good manner. By incorporating Huffman encoding in a search based chaotic cryptosystem, any organization can benefit from an online resource infrastructure with secure data.

### REFERENCES