An algorithm for selection of the multiplicity of correctable errors for the data protection on application of the Reed-Solomon codes

P A Rahman\(^1\) and V E Bolnokin\(^2\)

\(^1\)Department of Automated Technological and information Systems, Ufa State Petroleum Technological University, 2, October Avenue, Sterlitamak, 453118, Russian Federation

\(^2\)Mechanical Engineering Research Institute named A.A. Blagonravov of the Russian Academy of Sciences, Moscow, Russian Federation

E-mail: pavelar@yandex.ru, vitalybolnokin@yandex.ru

Abstract. Within the scope of this scientific paper the probability analysis of unrecoverable data corruption in the data transmission systems and data protection on application of the Reed-Solomon codes are discussed. An offered by the author algorithm for selection of the multiplicity of correctable errors for the given bit error probability, fixed size of data frames, minimum required size for the user data and maximum acceptable probability of unrecoverable data corruption on application of the Reed-Solomon codes is also given. Finally, results of the experimental research of the probability of unrecoverable data corruption and example of selection of the multiplicity of correctable errors are also provided.

1. Introduction

In modern world the information technologies are an essential part of human life and business processes of enterprises. However, because of imperfection of the data transmission and storage systems [1, 2] in which information can be corrupted under the influence of external noises or physical damages, the problem of data protection is quite urgent [3, 4].

At the present day several types of the information redundancy technologies are used in the modern data storage and transmission systems with the use of the special data encoding algorithms based on the error-correcting codes, in particular the Reed-Solomon codes [5-7], which allow correcting the errors in the corrupted data due to use of the redundant (control) information. However, use of the redundant information reduces the portion of user data in the transmitted through the network data frames. So here we have the question of selection between the information redundancy and probability of the unrecoverable data corruption [8, 9].

Within the research work in the field of reliability of data storage, processing and transmission systems [10-12], a scientific problem of the probability analysis of unrecoverable data corruption in the data transmission systems and development of the algorithm for selection of the multiplicity of correctable errors for the data protection on application of the Reed-Solomon codes was raised by the author. Accordingly, the author carried out the theoretical analysis and experimental research of the probability of unrecoverable data corruption. The author also offered an algorithm for selection of the
multiplicity of correctable errors for the given probability of bit error, fixed size of data frames, minimum required size for the user data and maximum acceptable probability of unrecoverable data corruption on application of the Reed-Solomon codes.

2. Data encoding on application of the Reed-Solomon codes

In modern practice of redundant coding on application of the Reed-Solomon codes [5-7] the user data are represented as the blocks of \( k \) bytes. The \( r \) redundant (control) bytes are calculated by using the encoding algorithm for the Reed-Solomon codes based on the algebraic representation of information in the form of polynomials over the Galois field \( \text{GF}(2^8) \) and their algebraic transformations over this field. Finally, in case of use of the systematic Reed-Solomon coding, the control bytes are added after the user bytes, and, as a result, a data frame of \( n \) bytes is formed (figure 1).

![Figure 1. Structure of the data frame with user and control bytes.](image)

Here, \( M_i \) is user bytes and \( R_j \) is control bytes.

The number of control bytes \( r \) is even, and it equals to double of the multiplicity of correctable errors \( t \). In accordance to the error-correcting codes theory \( r = 2t \) control bytes allow correcting with 100\% guarantee up to \( t \) corrupted bytes at any position of the data frame and for any values of errors by using the decoding algorithm for the Reed-Solomon codes.

Thus, size of data frames is \( n = k + r \) and number of the control bytes is \( r = 2t \). The maximum size of data frames is \( 2^8 - 1 = 255 \) bytes in case of use of the Reed-Solomon codes over the Galois field \( \text{GF}(2^8) \). Accordingly, for the given fixed size \( n \) of data frames an increase of number of the control bytes \( r \) with purpose to increase the multiplicity of correctable errors \( t \) and provide lower probability of the unrecoverable frame corruption leads to decrease of portion of the user data, and only \( k = n - 2t \) bytes will be available for the user data. So, here we have the question of selection between the information redundancy and tolerance to the data corruption and need to analyze the probability of unrecoverable corruption of the data frame on application of the Reed-Solomon codes.

3. Analysis of the probabilities of corruption of data frames

In modern transmission systems information is transmitted as a sequence of bits and each of them can be corrupted. In most cases the probability of bit error is known for the data transmission system. To simplify the probability analysis we will use the following assumptions:

- The probability of error in any bit within the data frame is fixed and does not depend on the position of the byte in the data frame and position of a bit in the byte.
- The data transmission channel does not have «memory» and the probability of error in the next transmitted bit does not depend on whether the previous bits were corrupted.
- The probability of bit error does not change with time or changes rather slowly, and, within the time interval required to transmit the data frame, it is possible to consider the probability of bit error as a constant value.

If all of the above conditions are met, then transmission of the data frame of size \( n \) can be considered as a sequence of \( n \) independent transmissions of separate bytes. In turn, transmission of the byte can be considered as a sequence of 8 independent transmissions of 8 separate bits. As a result, we have a sequence of \( 8n \) transmission operations, which from the viewpoint of the probability theory can be interpreted as the \( n \) independent series of 8 independent tests in each series. We also should note that in the case of corruptions of several bits, they can be located in various ways inside a frame consisting of bytes. For example, 8 corrupted bits can be located inside one byte, as well as they can
be located in different bytes (up to 8 bytes). It should be noted, that these cases will be considered as different situations from the viewpoint of the correcting ability of the Reed-Solomon codes.

To analyze the probabilities of corruption of one, several or all bytes in data frames for the transmission systems with the given probability of bit error we need to use the mathematical methods of the probability theory [8-9].

Let $p$ is the given probability of bit error. Then, the probability of byte corruption is equal to probability of that at least one bit in byte is corrupted and it can be calculated by following formula:

$$P_{byte} = 1 - (1 - p)^8.$$  \hspace{1cm} (1)

Next, by taking into consideration the assumptions mentioned above, we can use the binomial law of distribution for the random number $T$ of corrupted bytes and obtain the probability of that exactly $h$ bytes are corrupted in the data frame for the given size $n$ of frame and probability $p$ of bit error:

$$P(T = h) = C_n^h (1 - (1 - p)^8)^h (1 - p)^{(n-h)}.$$  \hspace{1cm} (2)

The obtained formula determines the probability of corruption of the exact number of bytes $h$ under condition that the rest of $n - h$ bytes are intact for all of the appropriate combinations of corruption of $q = h...8h$ bits under condition that the rest of $8n - q$ bits are intact and at least one corrupted bit is located inside the each of $h$ corrupted bytes. Also, the formula takes into consideration all the combinations of $h$ corrupted bytes among $n$ bytes in the data frame.

Now, by using the obtained formula for calculation of corruption probability for the exact number of bytes $h$, and the given size $n$ of the data frame and probability of bit error $p$, we can easily derive the formulas for the probability of no corruptions ($T = 0$) in the data frame, probability of the recoverable corruption ($1 \leq T \leq t$) and probability of the unrecoverable corruption ($T > t$) on application of the Reed-Solomon codes for the given multiplicity $t$ of correctable errors.

Accordingly, the formula for the probability of no corruptions in the data frame:

$$P(T = 0) = (1 - p)^{8n}.$$  \hspace{1cm} (3)

Next, formula for the probability of the recoverable corruptions in the data frame:

$$P(1 \leq T \leq t) = \sum_{h=1}^{t} C_n^h (1 - (1 - p)^8)^h (1 - p)^{(n-h)}.$$  \hspace{1cm} (4)

Finally, formula for the probability of the unrecoverable corruptions in the data frame:

$$P(T > t) = \sum_{h=t+1}^{n} C_n^h (1 - (1 - p)^8)^h (1 - p)^{(n-h)}.$$  \hspace{1cm} (5)

4. Algorithm for selection of the multiplicity of correctable errors
As it was discussed above, to encode the user data with ability of correction up to $t$ corrupted bytes on application of the Reed-Solomon codes we need to sacrifice $r = 2t$ bytes in the data frame. So, in case of the given fixed size $n$ of data frames only $k = n - 2t$ bytes are available for the user data.

Obviously, to formulate a task of selection of reasonable multiplicity of correctable errors we need some criteria and restrictions. Let us use the next source parameters for the task of selection of multiplicity of correctable errors:

- $n$ is the fixed size of data frames (in bytes).
- $p$ is the probability of bit error.
- $k_{\text{min}}$ is minimum required size for the user data (in bytes).
- $P_{\text{max}}$ is maximum acceptable probability of unrecoverable data corruption.

Accordingly, by taking into consideration that the probability of unrecoverable data corruption should be less or equal to $P_{\text{max}}$ and number of bytes $k = n - 2t$ available for the user data should be...
greater or equal to $k_{\text{min}}$, we can offer the next simple mathematical model for the task of selection of reasonable multiplicity of correctable errors $t^*$:

$$
\begin{align*}
P(T > t) & \leq P_{\text{max}}; \\
\frac{n - 2t}{n} & \geq k_{\text{min}}; \\
t & \rightarrow \min; \quad t \geq 0.
\end{align*}
$$

It should be noted, that if the probability of corruption of at least one byte $P(T > 0)$ is less or equal to the given maximum acceptable probability of unrecoverable data corruption, then we obtain $t^* = 0$, and it is obvious, that application of the Reed-Solomon codes is not reasonable.

Moreover, if $P(T > 0) > P_{\text{max}}$, then application of the Reed-Solomon codes is reasonable, however in some cases it is impossible to find the reasonable multiplicity of correctable errors $t^*$, which fulfils the both conditions in the mathematical model (6).

Figure 2 shows the algorithm for selection of the multiplicity of correctable errors on application of the Reed-Solomon codes.

Figure 2. Algorithm for selection of the multiplicity of correctable errors.
5. Example of selection of the multiplicity of correctable errors

Let the next parameters for selection of the multiplicity of correctable errors are given: size of data frames is \( n = 36 \) bytes, minimum required size for the user data is \( k_{\text{min}} = 24 \) bytes, probability of bit error is \( p = 5 \cdot 10^{-4} \) and maximum acceptable probability of unrecoverable corruption is \( P_{\text{max}} = 10^{-6} \).

At first, let us calculate the probability of corruption of at least one byte in the data frame by the formula 5:

\[
P(T > 0) = 0.134143.
\]

Obviously, \( P(T > 0) > P_{\text{max}} \), therefore application of the Reed-Solomon codes is reasonable.

Next, let find the reasonable multiplicity of correctable errors, which fulfills to the next conditions:

\[
\begin{align*}
P(T > t) &\leq 10^{-6}; \\
36 - 2t &\geq 24.
\end{align*}
\]

It is easy to see, that 6 variants of multiplicity \( t = 1 \ldots 6 \) are fulfill to the condition \( 36 - 2t \geq 24 \). We need to find the minimal \( t \), for which the probability of unrecoverable corruption \( P(T > t) \leq 10^{-6} \).

For the multiplicities \( t = 1, 2 \) and 3 we obtain the following probabilities of unrecoverable data corruption by the formula 5:

\[
\begin{align*}
P(T > 1) &= 0.009179082. \\
P(T > 2) &= 0.000411863. \\
P(T > 3) &= 0.000013520.
\end{align*}
\]

As we can see, the obtained probabilities are greater than \( 10^{-6} \).

Finally, for the multiplicity \( t = 4 \) we obtain the probability of unrecoverable data corruption:

\[
P(T > 4) = 0.000000345.
\]

Obviously, \( P(T > 4) \leq 10^{-6} \), therefore the reasonable multiplicity is \( t^* = 4 \).

6. Experimental research of the probability of unrecoverable corruption

To provide the experimental research of the probability of unrecoverable data corruption the author developed special software, which provides generation, encoding, corruption and decoding of data frames on application of the Reed-Solomon codes.

Figure 3 shows the main window of the developed software. The developed software allows to generate and encode the given number of data frames with the size of \( 1 \leq n \leq 255 \) bytes and multiplicity of correctable errors \( 1 \leq t \leq 127 \), corrupt the binomially distributed random number of bytes for the given probability of bit error \( p \), decode data frames and collect the information about the results of decoding on application of the Reed-Solomon codes.

There are five cases of the frame decoding:

- Frame is recognized as intact and it is equal to the generated frame. This is the case of no corruptions, when actual number of corrupted bytes is zero \( (T = 0) \).
- Frame is corrected and the corrected frame is equal to the generated frame. This is the case of recoverable corruptions, when the actual number of corrupted bytes does not exceed the multiplicity of correctable errors \( (1 \leq T \leq t) \).
- Frame is recognized by the decoder as uncorrectable. This is the special case of unrecoverable corruption, when the actual number of corrupted bytes greater than the multiplicity of correctable errors \( (T > t) \) and the decoder cannot correct the errors.
- Frame is corrected, but the corrected frame differs from the generated frame. This is the special case of unrecoverable corruption, when the actual number of corrupted bytes greater than the multiplicity of correctable errors \( (T > t) \) and the decoder makes wrong corrections.
Frame is recognized as intact, but it differs from the generated frame. This is the special case of unrecoverable corruption, when the actual number of corrupted bytes greater than the twice of the multiplicity of correctable errors \((T > 2t)\) and the decoder cannot detect the errors.

The probability of the case 1 \((T = 0)\) can be estimated by the formula 3 and probability of the case 2 \((1 \leq T \leq t)\) can be estimated by the formula 4. The probability of the cases 3, 4 and 5 in total \((T > t)\) can be estimated by the formula 5.

For the given above example of selection of the multiplicity of correctable errors the author carried out an experimental research of the probability of unrecoverable data corruption for the given probability of bit error \(p = 5 \cdot 10^{-4}\), size \(n = 36\) of data frames and several multiplicities of correctable errors \(t = 1, 2, 3\) and 4.

For each of the multiplicities \(t = 1, 2, 3\) and 4 the separate series of 100 experiments were carried out and in each of them 1000000000 frames were generated, encoded, corrupted and decoded by using the developed software and the average number of the unrecoverable frames (decoding cases 3, 4 and 5 in total) was measured. The results of the experimental research are shown in Table 1.

It is easy to see, that the obtained average numbers of the unrecoverable frames confirm the calculated above probabilities of the unrecoverable data corruption.

### Table 1. Average number of the unrecoverable frames.

| \(t\) | Number of the generated frames | Average number of the unrecoverable frames |
|------|-------------------------------|-------------------------------------------|
| 1    | 1000000000                    | 9179037                                   |
7. Conclusion
Thus, within the scope of this scientific research the author analyzed the probability of unrecoverable data corruption in the data transmission systems and offered an algorithm for selection of the multiplicity of correctable errors for the data protection on application of the Reed-Solomon codes.

The obtained results were used for development of the specialized software and laboratory works aimed to analyze the probabilities of recoverable and unrecoverable corruptions of data frames on application of the Reed-Solomon codes for the students of technical specialties.

Acknowledgements
The author is grateful to professor I.I. Ladygin from Moscow Power Engineering Institute for scientific support and theoretical base in the field of error-correcting codes.

References
[1] Olifer V G and Olifer N A 2016 Computer Networks (Saint-Petersburg: Piter)
[2] Khurshidov A S 2001 The Essential Guide to Computer Data Storage (Prentice Hall)
[3] Thomasian A and Blaum M 2009 ACM Transactions on Storage 5(3) 7
[4] Tsaregorodtsev A V, Kravets O Ja, Choporov O N and Zelenina A N 2018 International Journal on Information Technologies and Security 10(4) 67-76
[5] Moon T K 2005 Error correcting coding: mathematical methods and algorithms (New Jersey: John Wiley & Sons Inc.)
[6] Wicker S B and Bhargava V K 1999 Reed-Solomon codes and their application (Wiley-IEEE Press)
[7] Blahut R E 1983 Theory and Practice of Error Control Codes (Reading, MA: Addison-Wesley)
[8] Gut A 2005 Probability: A Graduate Course (New York: Springer-Verlag)
[9] Gnedenko B V 2005 Probability theory course (Moscow: Editorial URSS)
[10] Rahman P A 2017 IOP Conf. Ser.: Mater. Sci. Eng. 177 012087
[11] Rahman P A and Bobkova E Yu 2017 J. Phys.: Conf. Ser. 803 012124
[12] Rahman P A and Bobkova E Yu 2017 J. Phys.: Conf. Ser. 803 012125