Analysis of mean time to data loss of fault-tolerant disk arrays RAID-6 based on specialized Markov chain

P A Rahman and G D'K Novikova Freyre Shavier

Ufa State Petroleum Technological University, 2, October Avenue, Sterlitamak,
453118, Russia

E-mail: pavelar@yandex.ru

Abstract. This scientific paper is devoted to the analysis of the mean time to data loss of redundant disk arrays RAID-6 with alternation of data considering different failure rates of disks both in normal state of the disk array and in degraded and rebuild states, and also nonzero time of the disk replacement. The reliability model developed by the authors on the basis of the Markov chain and obtained calculation formula for estimation of the mean time to data loss (MTTDL) of the RAID-6 disk arrays are also presented. At last, the technique of estimation of the initial reliability parameters and examples of calculation of the MTTDL of the RAID-6 disk arrays for the different numbers of disks are also given.

1. Introduction

These days the data storage systems [1, 2] are widely used as hardware platform for the information systems, which provides the business processes in the modern enterprises. The stability of the information systems directly depends on reliability of the data storage systems. Therefore, to increase the fault-tolerance of the data storage systems the disk arrays with data striping are often applied.

The key reliability index of the modern disk arrays is the mean time to data loss, and, accordingly, development of the reliability models for the disk arrays is an actual scientific problem.

Nowadays there is a number of academic books [3-6] dedicated to the reliability theory, containing the simplified reliability models of technical systems, which do not consider specific features of the modern disk arrays with data striping and provide quite overestimated values of the mean time to data loss. Also there is a set of the specialized reliability models for the disk arrays [7-10], which do not take into account the time necessary for the faulty disks replacement.

Accordingly, within the scope of this scientific paper, the authors offered a reliability model taking into account the time of disks replacement for the RAID-6 disk arrays with data striping.

2. Data redundancy in the RAID-6 disk arrays

The RAID-6 disk array consists of $n \geq 4$ independent disks with equal capacity. It ensures safe operation and data availability at failure of maximum two disks (any of them). The effective capacity of the disk array is equal to $(n - 2) / n$ part of total capacity of all disks. The $2 / n$ part of disk space of each disk is intended for storage of redundant (control) data, which is calculated on the basis of user data stored on the other disks.

In case of failure of any one or two disks, it is possible to calculate missing information according to user and control data stored on the remaining disks.

In case of failure of three disks, as well as in case of failure of any third disk before replacement at least of one of the faulty disks and completion of the data replication process on the replaced disks, all data of entire array become irreparably lost.

So, one may consider the RAID-6 array as a good compromise between the fault-tolerance and redundancy. The following figure 1 gives the distribution scheme of user and redundant data blocks in the RAID-6 array with five disks as an example.

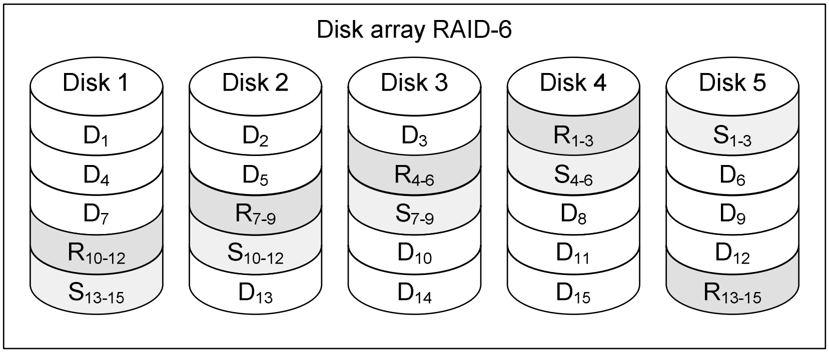

Figure 1. Distribution of the data blocks in the RAID-6 array with five disks.

The RAID-6 technology divides data into blocks and place them in “horizontal sections”, enclosing one block of every of *n* disks. In each section, there are allocated *n* – 2 information blocks $D_a \dots D_b$ and two control blocks S_{a-b} and R_{a-b}. Each block contains an array of bytes. The control bytes are allocated in the control blocks, and they are independently calculated according to the corresponding information bytes allocated in the information blocks.

Calculation of all the missing blocks in case of unavailability of any one or two disks is provided due to the control bytes in the S_{a-b} and R_{a-b} blocks.

3. Analysis of the mean time to data loss of disk arrays RAID-6

At first, let us consider a reliability model, based on Markov chain, offered by the authors for the RAID-6 disk arrays with data striping, taking into account the different rates of disks failure in normal state, degraded and rebuild states, and nonzero time of the faulty disk replacement. The reliability model uses the following set of states of the RAID-6 disk arrays and transitions between them:

- State 0 – normal state: all *n* disks of array are operable and data of the array are available. The array can pass from this state to state 1 with rate $n\lambda_0$ (failure of any operable disk).
- State 1 – degraded state 1: one disk is faulty and waits for replacement, the remaining *n* – 1 disks of the array are operable, data of the array are available. The array can pass from this state either to state 2 with rate $(n-1)\lambda_1$ (failure of one of the remaining disks) or to state 3 with rate μ_{D} (replacement of the faulty disk).
- State 2 – degraded state 2: two disks are faulty and wait for replacement, the remaining *n* – 2 disks of the array are operable, data of the array are available. The array can pass from this state either to state F with rate $(n-2)\lambda_2$ (failure of one of the remaining disks), or to state 4 with rate $2\mu_{\mathrm{D}}$ (replacement of one of the faulty disks).
- State 3 – rebuild state 3: the faulty disk is replaced and is involved in the data replication process, the remaining *n* – 1 disks of the array are operable, data of the array are available. The array can pass from this state either to state 0 with rate θ_1 (completion of the data replication on the replaced disk), or to state 1 with rate λ_{R} (failure of the replaced disk during the data replication), or to state 4 with rate $(n-1)\lambda_1$ (failure of one of the operable disks), or to state 5 with rate $(n-1)\varepsilon_1$ (read error on one of the operable disks).

- State 4 – rebuild state 4: one of the faulty disks is replaced and is involved in the data replication process, the other faulty disk waits for replacement, the remaining *n* – 2 disks of the array are operable, data of the array are available. The array can pass from this state either to state 1 with rate θ_2 (completion of the data replication on the replaced disk), or to state 2 with rate λ_R (failure of the replaced disk during the data replication), or to state 5 with rate μ_D (replacement of the second faulty disk), or to state F with rate $(n-2)\lambda_2$ (failure of one of the operable disks) or with rate $(n-2)\varepsilon_2$ (read error on one of the operable disks during the data replication).
- State 5 – rebuild state 5: the both faulty disks have been replaced and are involved in the data replication process, the other *n* – 2 disks of the array are operable, data of the array are available. The array can pass from this state either to state 0 with rate θ_2 (completion of the data replication on the both replaced disks), or to state 4 with rate $2\lambda_R$ (failure of one of the replaced disks during the data replication), or to state F with rate $(n-2)\lambda_2$ (failure of one of the operable disks) or with rate $(n-2)\varepsilon_2$ (read error on one of the operable disks during the data replication).
- State F – failed state: data of the disk array are unavailable and irreparably lost.

Accordingly, the Markov chain, which represents the discussed above set of states and transitions between the states, are shown below (figure 2):

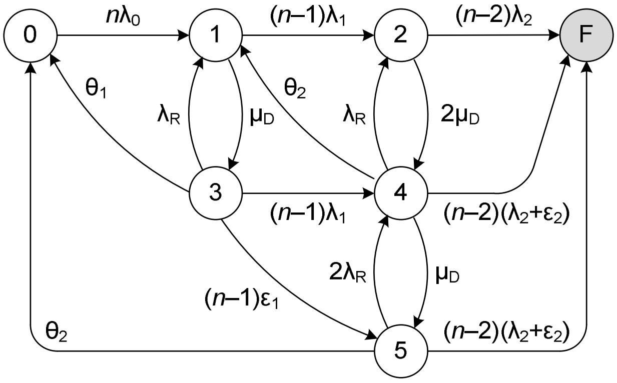

Figure 2. The specialized Markov chain for the RAID-6 disk arrays.

where, λ_0 – failure rate of disks in normal state of the disk array;

λ_1 – failure rate of disks in case of unavailability of one disk (extra load on the operable disks due to the data calculation of one unavailable disk causes a higher failure rate of the operable disks);

λ_2 – failure rate of disks in case of unavailability of two disks (extra load on the operable disks due to the data calculation of two unavailable disks causes a higher failure rate of the operable disks);

λ_R – failure rate of the replaced disk during the data replication (high amount of recorded operations on the replaced disk causes a significantly higher failure rate of the replaced disk);

μ_D – rate of replacement of the faulty disk;

θ_1 – rate of the data replication on the replaced disk in case of unavailability of one disk in the array;

θ_2– rate of the data replication on replaced disks in case of unavailability of two disks in the array;

ε_1 – rate of read errors on the operable disks during the data replication in case of unavailability of one disk in the array;

ε_2 – rate of read errors on the operable disks during the data replication in case of unavailability of two disks in the array.

Accordingly, the Kolmogorov-Chapman differential equations system for the Markov chain discussed above is as follows:

$$\begin{cases} P_0(0)=1; \quad P_1(0)=P_2(0)=P_3(0)=P_4(0)=P_5(0)=P_{\mathrm{F}}(0)=0; \\ P_0(t)+P_1(t)+P_2(t)+P_3(t)+P_4(t)+P_5(t)+P_{\mathrm{F}}(t)=1; \\ dP_0(t)/dt=-n\lambda_0 P_0(t)+\theta_1 P_3(t)+\theta_2 P_5(t); \\ dP_1(t)/dt=n\lambda_0 P_0(t)-(\mu_{\mathrm{D}}+(n-1)\lambda_1)P_1(t)+\lambda_{\mathrm{R}} P_3(t)+\theta_2 P_4(t); \\ dP_2(t)/dt=(n-1)\lambda_1 P_1(t)-(2\mu_{\mathrm{D}}+(n-2)\lambda_2)P_2(t)+\lambda_{\mathrm{R}} P_4(t); \\ dP_3(t)/dt=\mu_{\mathrm{D}} P_1(t)-(\theta_1+\lambda_{\mathrm{R}}+(n-1)(\lambda_1+\varepsilon_1))P_3(t); \\ dP_4(t)/dt=2\mu_{\mathrm{D}} P_2(t)+(n-1)\lambda_1 P_3(t)- \\ -(\mu_{\mathrm{D}}+\theta_2+\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))P_4(t)+2\lambda_{\mathrm{R}} P_5(t); \\ dP_5(t)/dt=(n-1)\varepsilon_1 P_3(t)+\mu_{\mathrm{D}} P_4(t)-(\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))P_5(t); \\ dP_{\mathrm{F}}(t)/dt=(n-2)\lambda_2 P_2(t)+(n-2)(\lambda_2+\varepsilon_2)P_4(t)+(n-2)(\lambda_2+\varepsilon_2)P_5(t). \end{cases} \tag{1}$$

Considering that the initial state of the RAID-6 disk array is state 0 and only in the states 0-5 the disk array is operable and user data are available, we can derive the formula for calculation of the mean time to data loss of the RAID-6 array considering it as the mean time of staying of the disk array in the states 0-5:

$$T_{\mathrm{R6DL}}=\int_0^\infty (P_0(t)+P_1(t)+P_2(t)+P_3(t)+P_4(t)+P_5(t))dt. \tag{2}$$

On the basis of advanced mathematical analysis, the authors solved the mathematical problem and obtained the formula for calculation of the mean time to data loss of the RAID-6 array:

$$T_{\mathrm{R6DL}}=\frac{W+M_1+M_2+M_3+M_4+M_5}{(n-2)\lambda_2 M_2+(n-2)(\lambda_2+\varepsilon_2)(M_4+M_5)}. \tag{3}$$

In this case:

$$\begin{gathered} W=\mu_{\mathrm{D}}(\theta_1+(n-1)(\lambda_1+\varepsilon_1))(\lambda_{\mathrm{R}}(n-2)\lambda_2(\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))+(2\mu_{\mathrm{D}}+(n-2)\lambda_2)\times \\ \times((\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))(n-2)(\lambda_2+\varepsilon_2)+\mu_{\mathrm{D}}(\theta_2+(n-2)(\lambda_2+\varepsilon_2))))+ \\ +\mu_{\mathrm{D}}\theta_2(2\mu_{\mathrm{D}}+(n-2)\lambda_2)((\theta_2+(n-2)(\lambda_2+\varepsilon_2))(\theta_1+(n-1)\varepsilon_1)+2\lambda_{\mathrm{R}}\theta_1)+ \\ +(n-1)\lambda_1(\theta_1+\lambda_{\mathrm{R}}+(n-1)(\lambda_1+\varepsilon_1))(\mu_{\mathrm{D}}(2\mu_{\mathrm{D}}+(n-2)\lambda_2)(\theta_2+(n-2)(\lambda_2+\varepsilon_2))+ \\ +(\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))((\theta_2+\lambda_{\mathrm{R}})(n-2)\lambda_2+(2\mu_{\mathrm{D}}+(n-2)\lambda_2)(n-2)(\lambda_2+\varepsilon_2))). \\ M_1=n\lambda_0(\theta_1+\lambda_{\mathrm{R}}+(n-1)(\lambda_1+\varepsilon_1))((2\mu_{\mathrm{D}}+(n-2)\lambda_2)(\mu_{\mathrm{D}}+\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))\times \\ \times(\theta_2+(n-2)(\lambda_2+\varepsilon_2))+\lambda_{\mathrm{R}}(n-2)\lambda_2(\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))). \\ M_2=n\lambda_0((n-1)\lambda_1(\lambda_{\mathrm{R}}(\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))\times \\ \times(\mu_{\mathrm{D}}+\theta_1+\lambda_{\mathrm{R}}+(n-1)(\lambda_1+\varepsilon_1))+(\theta_1+\lambda_{\mathrm{R}}+(n-1)(\lambda_1+\varepsilon_1))\times \\ \times(\mu_{\mathrm{D}}+\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))(\theta_2+(n-2)(\lambda_2+\varepsilon_2)))+2\lambda_{\mathrm{R}}^2\mu_{\mathrm{D}}(n-1)\varepsilon_1). \\ M_3=n\lambda_0\mu_{\mathrm{D}}((2\mu_{\mathrm{D}}+(n-2)\lambda_2)(\mu_{\mathrm{D}}+\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))\times \\ \times(\theta_2+(n-2)(\lambda_2+\varepsilon_2))+\lambda_{\mathrm{R}}(n-2)\lambda_2(\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))). \\ M_4=n\lambda_0\mu_{\mathrm{D}}((n-1)\lambda_1(\theta_2+2\lambda_{\mathrm{R}}+(n-2)(\lambda_2+\varepsilon_2))(2(\theta_1+\lambda_{\mathrm{R}}+(n-1)(\lambda_1+\varepsilon_1))+ \\ +(2\mu_{\mathrm{D}}+(n-2)\lambda_2))+2\lambda_{\mathrm{R}}(n-1)\varepsilon_1(2\mu_{\mathrm{D}}+(n-2)\lambda_2)). \\ M_5=n\lambda_0\mu_{\mathrm{D}}(2\mu_{\mathrm{D}}(n-1)\lambda_1(\theta_1+\lambda_{\mathrm{R}}+(n-1)(\lambda_1+\varepsilon_1))+(2\mu_{\mathrm{D}}+(n-2)\lambda_2)\times \\ \times(\mu_{\mathrm{D}}(n-1)\lambda_1+(n-1)\varepsilon_1(\mu_{\mathrm{D}}+\theta_2+(n-2)(\lambda_2+\varepsilon_2)))+\lambda_{\mathrm{R}}(n-1)\varepsilon_1(n-2)\lambda_2). \end{gathered}$$

Note 1. In case when the faulty disk replacement rate is $\mu_D \to \infty$ (average replacement time for the faulty disks tends to zero), the mathematical analysis provides the following simplified formula:

$$T_{\mathrm{R6DL}} = \frac{(\theta_1 + n\lambda_0 + (n-1)(\lambda_1 + \varepsilon_1))(\theta_2 + (n-2)(\lambda_2 + \varepsilon_2)) + n\lambda_0(n-1)(\lambda_1 + \varepsilon_1)}{n\lambda_0(n-1)(\lambda_1 + \varepsilon_1)(n-2)(\lambda_2 + \varepsilon_2)}. \quad (4)$$

Note 2. In case when the faulty disk replacement rate is $\mu_D = 0$ (no replacement of the faulty disks), the mathematical analysis provides the following simplified formula:

$$T_{\mathrm{R6DL}} = \frac{1}{n\lambda_0} + \frac{1}{(n-1)\lambda_1} + \frac{1}{(n-2)\lambda_2}. \quad (5)$$

4. Estimation of the initial reliability parameters for the RAID-6 disk arrays

The failure rate of disks λ_0 in operable state of the disk array can be estimated on the base of mean time to failure of disk T_{DF} (Mean Time to Failure), obtained from the practical experience or provided by the disk manufacturer. The failure rate of disks λ_1 and λ_2 in case of unavailability of one and two disks in the RAID-6 disk arrays is higher than the rate of λ_0 due to the fact that beside the primary load, the operable disks bears extra reading operations for calculation of data of the unavailable disks. One may simply consider that failure rate λ_1 is twice as high and failure rate λ_2 is three times as high. As for the failure rate of existing replicated disk λ_R, it is significantly higher than the λ_0 rate because of the large amount of the write operations. One may consider that λ_R is five times as high. Finally, taking into consideration the aforesaid, one can estimate parameters λ_0, λ_1, λ_2 and λ_R using the following simple formulas:

$$\lambda_0 = 1/T_{\mathrm{DF}}; \quad \lambda_1 = 2/T_{\mathrm{DF}}; \quad \lambda_2 = 3/T_{\mathrm{DF}}; \quad \lambda_R = 5/T_{\mathrm{DF}}. \quad (6)$$

The rate of disk replacement varies depending on the replacement method: whether the disk is replaced automatically using the additional spare disks and the hot-spare technology, or detection and disk replacement is manually carried out by technical specialists. However, in the both cases it is possible to conclude that the replacement rate is defined by the given (or obtained from the practical experience) mean time of waiting for spare (faulty disk replacement) τ_{WS}:

$$\mu_D = 1/\tau_{\mathrm{WS}}. \quad (7)$$

The rate of data replication θ_1 in case of unavailability of one disk depends upon given capacity V (byte) of disks, average recorded speed v_{WR} (byte/s) for disks and average data recalculation speed v_{C1} (byte/s) for the disk array in case of unavailability of one disk. As for the case of two disks unavailability, data recalculation speed v_{C2} is obviously lower because the data recalculation takes more time, and therefore the rate of data replication θ_2 will be also lower. One can simply estimate rates θ_1 and θ_2 using the following formulas:

$$\theta_1 = \frac{3600\, v_{\mathrm{C1}} v_{\mathrm{WR}}}{V(v_{\mathrm{C1}} + v_{\mathrm{WR}})}; \quad \theta_2 = \frac{3600\, v_{\mathrm{C2}} v_{\mathrm{WR}}}{V(v_{\mathrm{C2}} + v_{\mathrm{WR}})}. \quad (8)$$

Estimation of the rate of data read errors ε_1 on operable disks during the data replication in case of unavailability of one disk is based on the probability of bit unrecoverable read error P_{URE}, provided by disks manufacturer or obtained from the practical experience, given disk capacity V (bytes) and calculated rate of data replication θ_1 (hour^{-1}). Similarly, one can estimate the rate of data read error ε_2 on the operable disks during the data replication process with rate θ_2 (hour^{-1}) in case of unavailability of two disks. One can estimate rates ε_1 and ε_2 using the following formulas:

$$\varepsilon_1 = 8V\theta_1 P_{\mathrm{URE}}; \quad \varepsilon_2 = 8V\theta_2 P_{\mathrm{URE}}. \quad (9)$$

5. Calculation examples of the mean time to data loss for the RAID-6 disk arrays

A set of *n* identical disks with capacity of $V_{\mathrm{D}} = 10^{12}$ bytes is given. The mean time to failure of disk is $T_{\mathrm{DF}} = 120000$ hours. The probability of bit unrecoverable read error is $P_{\mathrm{URE}} = 10^{-14}$. The average data recorded speed is $\nu_{\mathrm{WR}} = 50 \cdot 10^{6}$ byte/s.

A disk controller supporting the RAID-6 disk arrays is also given. The average data recalculation speed in case of unavailability of one disk in the disk array is $\nu_{\mathrm{C1}} = 15 \cdot 10^{6}$ byte/s. The average data recalculation speed in case of unavailability of two disks in the disk array is $\nu_{\mathrm{C2}} = 6 \cdot 10^{6}$ byte/s.

The mean time of waiting for spare disk (faulty disk replacement) is $\tau_{\mathrm{WS}} = 8$ hours.

At first, let us estimate the initial reliability parameters necessary for calculation of the mean time to data loss in the reliability model of the RAID-6 disk arrays, using formulas (6-9):

The failure rate of disks in normal state of disk array is: $\lambda_0 = 1/120000$ hour^{-1}.

The failure rate of disks in case of unavailability of one disk is: $\lambda_1 = 2/120000$ hour^{-1}.

The failure rate of disks in case of unavailability of two disks is: $\lambda_2 = 3/120000$ hour^{-1}.

The failure rate of the being replicated disk is: $\lambda_{\mathrm{R}} = 5/120000$ hour^{-1}.

The rate of faulty disks replacement is: $\mu_{\mathrm{D}} = 1/8$ hour^{-1}.

The rate of data replication in case of unavailability of one disk is: $\theta_1 \approx 1/24$ hour^{-1}.

The rate of data replication in case of unavailability of two disks is: $\theta_2 \approx 1/52$ hour^{-1}.

The rate of data read error on operable disks during the data replication process in case of unavailability of one disk is: $\varepsilon_1 = 1/300$ hour^{-1}.

The rate of data read error on operable disks during the data replication process in case of unavailability of two disks is: $\varepsilon_2 = 1/650$ hour^{-1}.

Now, let us calculate the mean time to data loss of the RAID-6 arrays for the different number of disks *n* = 4…12, using all estimated initial reliability parameters and the formula (3) discussed above for the RAID-6 disk arrays, and compare the MTTDL of the RAID-6 arrays with the MTTF of a single disk, which is given and equal to 120000 hours. The calculation results are shown in table 1.

Table 1. Results of calculation of mean time to data loss of the RAID-6 arrays.

n	Disk overhead, 2/*n* (%)	MTTDL, T_{R6DL} (hours)	$T_{\mathrm{R6DL}} / T_{\mathrm{DF}}$
4	50.00 %	1103005	9.192
5	40.00 %	502759	4.189
6	33.33 %	284173	2.368
7	28.57 %	182275	1.519
8	25.00 %	127074	1.059
9	22.22 %	93964	0.783
10	20.00 %	72584	0.605
11	18.18 %	57985	0.483
12	16.67 %	47570	0.396

6. Conclusion

The results of calculations show that growth of disks quantity leads to the inevitable decrease of the mean time to data loss of the RAID-6 disk arrays. Moreover, before reaching a certain quantity of disks, the RAID-6 disk arrays have a better value of the mean time to data loss than the mean time to failure of a single disk, and for the larger number of disks (in the example discussed above starting from 9 disks), the MTTDL of the RAID-6 array is worse.

Thus, the RAID-6 disk arrays with data striping are a good compromise between redundancy and fault-tolerance for a relatively small number of disks.

Acknowledgements

The authors are grateful to professor I. I. Ladygin from Moscow Power Engineering Institute for scientific support and theoretical base in the field of reliability models of technical systems.

References

[1] Nelson S 2011 *Pro data backup and recovery* (Apress) p 296
[2] Thomasian A and Blaum M 2009 *ACM Transactions on Storage* **5(3)** 7
[3] Shooman M L 2002 *Reliability of computer systems and networks* (John Wiley & Sons) p 528
[4] Koren I and Krishna C M 2007 *Fault-Tolerant Systems* (Morgan Kaufmann Publishers) p 378
[5] Cherkesov G N 2005 *Reliability of Hardware-Software Systems* (Saint-Petersburg: Piter) p 479
[6] Polovko A M and Gurov S V 2006 *Basis of Reliability Theory* (Saint-Petersburg: BHV-Petersburg) p 704
[7] Schwarz T J E and Burkhard W A 1995 *IEEE Transactions on Magnetics* **31(2)** 1161-1166
[8] Rahman P A 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **177** 012087
[9] Elerath J G and Pecht M 2009 *IEEE Transactions on Computers* **58(3)** 289-299
[10] Rahman P A 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **177** 012088