Reliability model of disk arrays RAID-5 with data striping

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. Within the scope of the this scientific paper, the simplified reliability model of disk arrays RAID-5 (redundant arrays of inexpensive disks) and an advanced reliability model offered by the authors taking into the consideration nonzero time of the faulty disk replacement and different failure rates of disks in normal state of the disk array and in degraded and rebuild states are discussed. The formula obtained by the authors for calculation of the mean time to data loss (MTTDL) of the RAID-5 disk arrays on basis of the advanced model is also presented. Finally, the technique of estimation of the initial reliability parameters, which are used in the reliability model, and the calculation examples of the mean time to data loss of the RAID-5 disk arrays for the different number of disks are also given.

1. Introduction

In present days most of the modern enterprises use different kind of the data storage systems [1, 2] for the files and databases, which are used within the enterprise business processes. The availability and integrity of the files and databases directly depend on reliability of the data storage systems.

In the modern computer world, the disk arrays with data striping are often used as an essential part of the data storage systems. Therefore, the development of the reliability models for the disk arrays and analysis of such important reliability index as the mean time to data loss are a quite urgent task.

The modern literature on the reliability theory [3-6] offers a number of the simplified reliability models of the technical systems, based on the Markov birth-death chains. Unfortunately, these models do not consider specific aspects of functioning of the disk arrays with data striping and give overestimated values for the mean time to data loss. Also there is a set of scientific papers [7-10], which offer a set of the specialized reliability models for the disk arrays; however, they do not consider the time of the faulty disk replacement.

Respectively, within this scientific paper, the authors offered a specialized reliability model for the RAID-5 disk arrays with data striping taking into consideration the disk replacement time.

2. Structure of the RAID-5 disk arrays

The RAID-5 disk arrays consist of $n \geq 3$ independent disks with identical capacity and remain operable in case of failure of no more than one (any) disk.

The effective capacity of the RAID-5 array is $(n - 1) / n$ part of the total capacity of disks. The $1 / n$ part of each disk space is intended for storage of the redundant (control) information calculated according to user data, which are located on the other disks. It allows one, in case of any single disk failure, to calculate missing information by using the user and control data on the remaining $n - 1$ disks.

In case of failure of any two disks, as well as in case of failure of any second disk prior to replacement of faulty disk and completion of the data replication process on replaced disk, all data of the entire array will be irreparably lost.

Thus, the RAID-5 disk arrays present a compromise between fault-tolerance and redundancy.

Figure 1 shows the distribution scheme of user and redundant data blocks by example of the RAID-5 array with four disks.

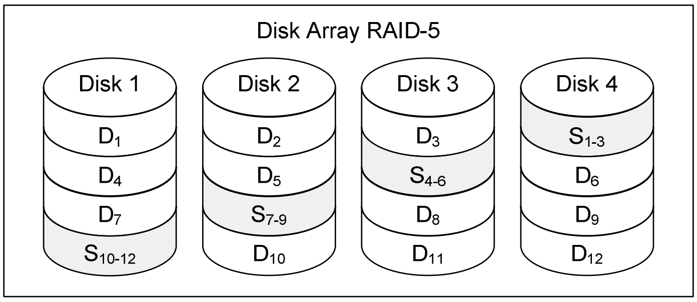

Figure 1. A structure of the RAID-5 disk array with four disks.

The RAID-5 technology divides data into blocks and places them in “horizontal sections” enveloping one block of each of *n* disks.

Each section contains *n* – 1 information blocks - $D_a \dots D_b$ and one control block - S_{a-b}. In turn, the blocks contain arrays of bytes. The control bytes are located in control block S_{a-b}, which are calculated on the base of appropriate information bytes located in the information blocks.

Due to control bytes in S_{a-b} blocks, it is possible to compute all missing blocks in case of unavailability of any one disk.

3. Simplified reliability model of RAID-5 disk arrays

In the simplified reliability model, on basis of the Markov birth-death chain [3-6], the RAID-5 disk array is considered as a system with *n* independent elements, which is tolerant to the failure of any single element. The initial state of the disk array is the state when all *n* disks are operable.

In the simplified reliability model of the RAID-5 disk array, the following set of states and transitions between the states is used:

- State 0 – normal state: all *n* disks of the array are operable and data of the array are available. From this state, the disk array can pass to state 1 with rate $n\lambda_D$ (failure of any disk).
- State 1 – degraded state: one of the disks in array fails, the remaining *n* – 1 disks of the array are operable, data of the array are available. From this state, the array can pass either to state 0 with rate θ_R (replacement of the failed disk and completion of the data replication on the replaced disk), or to state F with rate $(n-1)\lambda_D$ (failure of one of the operable disks).
- State F – failed state: data of the disk array are unavailable and irreparably lost.

Now let us present the following Markov chain considering the set of states discussed above and transitions between them (fig. 2):

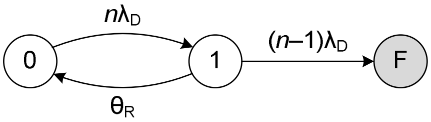

Figure 2. A state graph in the simplified reliability model of the RAID-5 disk arrays.

Here, λ_D – failure rate of the disks.

θ_R – rate of the data replication.

Accordingly, the Kolmogorov-Chapman differential equations system for the Markov chain is:

$$\begin{cases} P_0(0) = 1; \quad P_1(0) = P_\mathrm{F}(0) = 0; \\ P_0(t) + P_1(t) + P_\mathrm{F}(t) = 1; \\ dP_0(t)/dt = -n\lambda_\mathrm{D} P_0(t) + \theta_\mathrm{R} P_1(t); \\ dP_1(t)/dt = n\lambda_\mathrm{D} P_0(t) - (\theta_\mathrm{R} + (n-1)\lambda_\mathrm{D}) P_1(t); \\ dP_\mathrm{F}(t)/dt = (n-1)\lambda_\mathrm{D} P_1(t). \end{cases} \tag{1}$$

Considering that in the simplified reliability model only in states 0-1 of the RAID-5 disk array is operable and user data are available, one can derive the formula for calculation of the mean time to data loss of the RAID-5 array considering it as the mean time of preserving the disk array in the states 0-1 and taking into account that the initial state of the disk array is state 0:

$$T_\mathrm{R5DL} = \int_0^\infty (P_0(t) + P_1(t))dt. \tag{2}$$

Finally, one can obtain the following simplified formula for calculation of the mean time to data loss of the RAID-5 array using the mathematical analysis:

$$T_\mathrm{R5DL} = \frac{\theta_\mathrm{R} + (2n-1)\lambda_\mathrm{D}}{n(n-1)\lambda_\mathrm{D}^2}. \tag{3}$$

Mention must be made that in the simplified reliability model, the following important assumptions and simplifications are used:

- The disk failure rates are same for the normal and degraded states of the disk array.
- The time of replacement of the failed disk is considered as negligible (several seconds in case of using of the hot-spare disks) in comparison with the data replication (dozens of hours).
- The probability of the read errors on the remaining *n* – 1 disks during the data replication on the replaced disk and subsequent failure of the data replication process is ignored.

4. Advanced reliability model of RAID-5 disk arrays

Now let us overview the advanced reliability model on the basis of the specialized Markov chain offered by the authors for the RAID-5 disk arrays with data striping taking into consideration the nonzero time of disks replacement and different failure rate of disks in normal state and in degraded and rebuild states. In the advanced reliability model of the RAID-5 disk array, the following set of states and transitions between the states is used:

- State 0 – normal state: all *n* disks of the array are operable and data of the array are available. From this state, the disk array can pass to state 1 with rate $n\lambda_0$ (failure of any disk).
- State 1 – degraded state: one of the disks in the array fails and waits for replacement, the remaining *n* – 1 disks of the array are operable, data of the array are available. From this state, disk array can pass either to state F with rate $(n-1)\lambda_1$ (failure of one of the operable disks), or to state 2 with rate μ_D (replacement of the faulty disk).
- State 2 – rebuild state: the failed disk is replaced and the data replication process is started, the remaining *n* – 1 disks of the array are operable, data of the array are available. From this state, the array can pass either to state 0 with rate θ_1 (successful completion of data replication on the replaced disk), or to state 1 with rate λ_R (failure of the replaced disk during the data replication process), or to state F either with rate $(n-1)\lambda_1$ (failure of one of the operable disks) or with rate $(n-1)\varepsilon_1$ (read error on one of the operable disks during the data replication process).
- State F – failed state: data of the disk array are unavailable and irreparably lost.

Now let us present the following Markov chain considering the set of states and transitions between them discussed above (fig. 3):

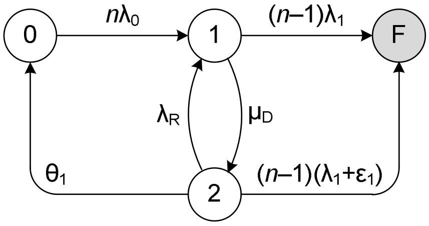

Figure 3. A state graph in the advanced reliability model of the RAID-5 disk arrays.

In the figure:

λ_0 – failure rate of disks in the normal state of the disk array.

λ_1 – failure rate of disks in case of unavailability of one disk (extra load on operable disks due to need of data calculation of the unavailable disk increases the 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 significantly increases the 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.

ε_1 – rate of read errors on the operable disks during the data replication on the replaced disk.

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

$$\begin{cases} P_0(0)=1; \quad P_1(0)=P_2(0)=P_F(0)=0; \\ P_0(t)+P_1(t)+P_2(t)+P_F(t)=1; \\ dP_0(t)/dt=-n\lambda_0P_0(t)+\theta_1P_2(t); \\ dP_1(t)/dt=n\lambda_0P_0(t)-(\mu_D+(n-1)\lambda_1)P_1(t)+\lambda_RP_2(t); \\ dP_2(t)/dt=\mu_DP_1(t)-(\theta_1+\lambda_R+(n-1)(\lambda_1+\varepsilon_1))P_2(t); \\ dP_F(t)/dt=(n-1)\lambda_1P_1(t)+(n-1)(\lambda_1+\varepsilon_1)P_2(t). \end{cases} \tag{4}$$

Considering that in the advanced reliability model, only in states 0-2, the RAID-5 disk array is operable and user data are available, one can derive the formula for calculation of the mean time to data loss of the RAID-5 array, considering it as the mean time of staying of the disk array in states 0-2 and taking into account that the initial state of the disk array is state 0:

$$T_{\mathrm{R5DL}}=\int_0^\infty (P_0(t)+P_1(t)+P_2(t))dt. \tag{5}$$

The authors solved the mathematical problem and obtained the advanced formula for calculation of the mean time to data loss of the RAID-5 array using the mathematical analysis:

$$T_{\mathrm{R5DL}}=\frac{(\mu_D+n\lambda_0+(n-1)\lambda_1)(\theta_1+(n-1)\varepsilon_1)+(n\lambda_0+(n-1)\lambda_1)(\mu_D+\lambda_R+(n-1)\lambda_1)}{n\lambda_0((n-1)\lambda_1(\theta_1+\lambda_R)+(\mu_D+(n-1)\lambda_1)(n-1)(\lambda_1+\varepsilon_1))}. \tag{6}$$

Note 1. If the faulty disk replacement rate is $\mu_D\to\infty$ (average replacement time for the faulty disks tends to zero), then the calculation formula is simplified to the following form:

$$T_{\mathrm{R5DL}}=\frac{\theta_1+n\lambda_0+(n-1)(\lambda_1+\varepsilon_1)}{n\lambda_0(n-1)(\lambda_1+\varepsilon_1)}. \tag{7}$$

Note 2. If the faulty disk replacement rate is $\mu_D=0$ (no replacement of the faulty disks), then the calculation formula is simplified to the following form:

$$T_{\mathrm{R5DL}}=\frac{1}{n\lambda_0}+\frac{1}{(n-1)\lambda_1}. \tag{8}$$

5. Estimation of the initial reliability parameters for the RAID-5 disk arrays

One can estimate the failure rate of disks λ_0 in normal state of the disk array on the basis of mean time to disk failure T_{DF} (Mean Time To Failure), either given by the disk manufacturer or retrieved from the practical experience. The failure rate of disks λ_1 in case of unavailability of any, disk data in the RAID-5 array exceed failure rate λ_0 due to the additional load on operable disks because of the extra read operations, required for calculation of data of the unavailable disk. One can simply suppose that in the worst case, failure rate λ_1 is twice as high. As for failure rate λ_{R} of the being replicated disk because of the large amount of recorded operations the failure rate of disk is significantly exceeds failure rate λ_0. In the worst case, one can simply suppose that failure rate λ_{R} is five times higher.

Finally, taking into account the aforesaid, for estimation of reliability parameters λ_0, λ_1 and λ_{R}, one can use following simple formulas:

$$\lambda_0 = 1/T_{\mathrm{DF}}; \quad \lambda_1 = 2/T_{\mathrm{DF}}; \quad \lambda_{\mathrm{R}} = 5/T_{\mathrm{DF}}. \tag{9}$$

The rate of replacement of the faulty disk varies depending on its method: whether replacement occurs automatically due to using the extra spare disks (in addition to the host drives in array) and hot-spare technology, or the faulty disk detection and replacement are manually carried out by the computer technician. In the first case, replacement of the faulty disk can take several minutes, in the second one, duration of this process can reach several hours. However, in the both cases one may say that the replacement rate can be estimated on the basis of the given (or retrieved from the practical experience) mean time of waiting for spare disk (replacement of the faulty disk) τ_{WS}:

$$\mu_{\mathrm{D}} = 1/\tau_{\mathrm{WS}}. \tag{10}$$

The rate of data replication θ_1 in case of unavailability of one disk in the RAID-5 disks arrays depends on the given V capacity (bytes) of disks, average data recorded speed v_{WR} (byte/s) of disks and average time of data recalculation speed v_{C1} (byte/s) in case of unavailability of one disk. The rate of data replication θ_1 can be estimated using the following formula:

$$\theta_1 = \frac{3600\, v_{\mathrm{C1}} v_{\mathrm{WR}}}{V(v_{\mathrm{C1}} + v_{\mathrm{WR}})}. \tag{11}$$

The rate of read errors ε_1 on the operable disks during the data replication in case of unavailability of one disk in the disk array can be estimated on the basis of the probability of bit unrecoverable read error P_{URE}, provided by the disk manufacturer or retrieved from the practical experience, given V capacity (bytes) of disks and the calculated rate of data replication θ_1 (hour^{-1}):

$$\varepsilon_1 = 8V\theta_1 P_{\mathrm{URE}}. \tag{12}$$

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

A set of n identical disks with capacity $V_{\mathrm{D}} = 10^{12}$ 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 of disks is $\nu_{\mathrm{WR}} = 50 \cdot 10^6$ byte/s.

A disk controller supporting the RAID-5 disk arrays is also given. The average data recalculation speed in case of unavailability of one disk in RAID-5 disk arrays is $\nu_{\mathrm{C1}} = 15 \cdot 10^6$ byte/s.

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

At first, let us estimate initial reliability parameters necessary for calculation of the mean time to data loss in the reliability model of the RAID-5 disks arrays, using formulas (9-12):

The failure rate of disks in the normal state of the 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 the being replicated disk is: $\lambda_R = 5/120000$ hour^{-1}.

The rate of replacement of the faulty disk is: $\mu_D = 1/8$ hour^{-1}.

The rate of the data replication on the replaced disk in the disk array is: $\theta_1 \approx 1/24$ hour^{-1}.

The rate of data read errors on the operable disks during the data replication is: $\varepsilon_1 = 1/300$ hour^{-1}.

Now, using all initial reliability parameters and advanced formula (6) for the RAID-5 disk arrays, let us calculate the mean time to data loss for the different number of disks *n* = 3…10, and compare the MTTDL of the RAID-5 arrays with the MTTF of a single disk, which is given and equal to 120000 hours. The calculation results are given in table 1.

Table 1. Results of calculations of mean time to data loss of the RAID-5 array.

n	Disk overhead, 1/*n* (%)	MTTDL,T_{R5DL} (hours)	T_{R5DL} / T_{DF}
3	33.33 %	288484	2.404
4	25.00 %	154262	1.285
5	20.00 %	98570	0.821
6	16.67 %	69273	0.577
7	14.29 %	52666	0.439
8	12.50 %	41648	0.347
9	11.11 %	34064	0.284
10	10.00 %	28588	0.238

7. Conclusion

The results of calculations show that the increase of disks quantity leads to a rapid decrease of mean time to data loss of the RAID-5 disk arrays.

Moreover, beginning from a certain number of disks in the RAID-5 array (beginning from 5 disks in the example overviewed above), the array has the worse value of the mean time to data loss than the mean time to failure of a single disk. Thus, the RAID-5 disk arrays with data striping represent an acceptable compromise between fault-tolerance and redundancy for the very small number of disks.

Finally, if one uses simplified formula (3) to calculate for example the mean time to data loss of the RAID-5 array for the number of disks *n* = 4, disk failure rate $\lambda_D = 1/120000$ hour^{-1} and data replication rate $\theta_R = 1/24$ hour^{-1}, then one will obtain a significantly overestimated value for the mean time to data loss: T_{R5DL} = 50070000 hours. Thus, the advanced formula (6) obtained by the authors for estimation of the mean time to data loss of the RAID-5 arrays provides more realistic values.

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] Khurshidov A S 2001 *The Essential Guide to Computer Data Storage* (Prentice Hall) p 317
[2] Thomasian A and Blaum M 2009 *ACM Transactions on Storage* **5(3)** 7
[3] Abd-El-Barr M 2007 *Design and Analysis of Reliable and Fault-Tolerant Computer Systems* (London: Imperial College Press) p 440
[4] Rausand M and Holyand A 2009 *System Reliability Theory* (John Wiley & Sons) p 643
[5] Ushakov I A 2008 *Reliability Theory Course* (Moscow: Drofa) p 239
[6] Ostreykovsky V A 2003 *Reliability Theory* (Moscow: Vyshaya Shkola) p 463
[7] Rahman P A 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **177** 012087
[8] Rahman P A 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **177** 012088
[9] Schwarz T J E and Burkhard W A 1995 *IEEE Transactions on Magnetics* **31(2)** 1161-1166

[10] Elerath J G and Pecht M 2009 *IEEE Transactions on Computers* **58(3)** 289-299