Mathematical Models of Reliability, Performance and Cost of an All-Flash Storage

Dmitry Kositsyn, Anton Shabaev,
Eugene Pitukhin, Vadim Ponomarev
Petrozavodsk State University
Petrozavodsk, Russia
kositsyn@psu.karelia.ru,
{ashabaev,eugene}@petrsu.ru,
vadim@cs.karelia.ru

Evgeny Ivashko
Institute of Applied Mathematical Research, KRC of RAS
Petrozavodsk State University
Petrozavodsk, Russia
ivashko@krc.karelia.ru

***Abstract*—The recent years there is rise of all-flash data storage. Flash disk has significant advantages comparing to HDD. This motivates to migrate existing and develop new high load systems basing on all-flash storage. Thus, one should choose an appropriate storage architecture to balance cost, throughput, load, performance, etc. In this paper we present mathematical models which describe reliability, performance and cost of all-flash data storage. Also, we provide results of simulations, analysis and recommendations for various usage scenarios. The results of this study support development of smart sensors for Internet of Things, where all-flash data storage is used to maintain sensed data locally.**

I. Introduction

There is growing demand for volumes of data storage services, at the same time performance requirements for storage systems are increasing [1]. The modern and future technologies of flash disks are the most promising for the future storage systems.

Comparing to so called “traditional” hard disk drives (HDD) a flash disk (or solid-state drive, SSD) has a number of advantages and disadvantages, but the former significantly overweight the latter. In particular, comparing to HDD, performance (access time, read and write throughput) of a flash disk is 1-2 orders of magnitude better, while the costs are an order of magnitude worse. Meanwhile, one of the most annoying a flash disk problem is so called “wear-out”. It is the result of a semiconductor memory cell damage during data write operations. Example of real-world SSD wear-out is shown in [2]. In the area of flash storage data deduplication is widely used to reduce both write wear-out and cost of storage (see for details [3] and [4]). This emerging technology is a relatively new method aimed to reduce redundancy by eliminating duplicate copies of data. The most prominent deduplication solution for Linux-based systems is Virtual Data Optimizer (VDO) [5].

A data storage system consists of special hardware assembling data disks. A common data storage is under high load of input/output operations. This affect on reliability (as failure of a disk) and performance. There is a well-known method to increase both reliability and performance of a data storage, called “RAID” (Redundant Array of Independent Disks), in which several physical drives are combined into one logical drive.

This paper presents an approach to analyze three characteristics of an all-flash storage: reliability, performance and cost of RAID. In contrast to our previous works, this paper considers the results of numerical modeling. The research results can be used to select the type of RAID for continuous smart data collection [6]. In particular, development of smart sensor systems for Internet of Things can use all-flash data storage to maintain sensed data locally. Note that such data can come from multiple sources, and the storage becomes subject to high requirements on reliability, performance and cost.

The structure of paper is the following. Section II describes a conceptual and three mathematical models: performance, cost and reliability. Section III presents the simulation results. Finally, section IV summarizes the final remarks and conclusions.

II. Data storage system models

Mathematical models are used to describe a system using mathematical concepts and language. Mathematical models are widely used to explain a system and to study the effects of different components, and to make predictions about behaviour. In this section a conceptual and three mathematical models are presented. The latter are used to describe performance, reliability and cost of an all-flash data storage.

A. Conceptual model

A conceptual model presented below is used to describe an all-flash data storage in general (see figure 1). It consists of external factors such as DSS control and an application input/output flows, and internal subsystems:

- network service;
- VDO-based deduplication software module;
- software RAID;
- disks subsystem consisting of a number of flash disks.

The purpose of a storage is to store and provide access to user data meeting the requirements of a service level agreement (SLA) in terms of reliability, performance, time and volume of data.

ISSN 2305-7254

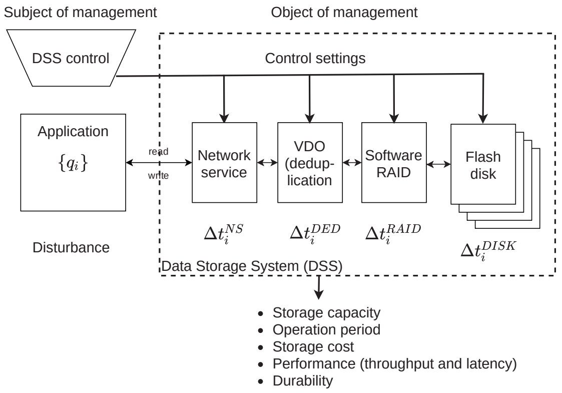


Fig. 1. Conceptual model of a data storage system

As input of a storage, request q_i ($i = 1, 2, ...$) is processed by the network service, which takes Δt_i^{NS} time to complete. From the point of view of our model, the primary function of the network service is to split the request into blocks of a certain size.

Passing the network service, blocks of the request are directed as input of the data deduplication system. As it was stated above, we consider VDO as the deduplication software module; its purpose is to detect and reduce data redundancy, for example, by replacing repeated copies of data with links to the first copy (see [3]). We denote the time required for deduplication of the request as Δt_i^{DED}.

The next module meeting the request is software RAID, which controls the process of reading or writing to underlying flash disks. The execution time of the request is denoted as Δt_i^{RAID}; it depends on the RAID level, type of operation and other minor factors.

The final point for the request is a flash disks subsystem. The access time Δt_i^{DISK} depends on the type of operation (read, write or erase) and the storage parameters, such as the average number of requests per second (IOPS, Input-output Operations Per Second) and average read or write throughput, measured in bits per second.

The bottom part of the Fig. 1 shows the output integral indicators that characterize the quality of storage systems and the efficiency of its operation for the user:

- storage capacity, bytes;
- operation period, hours;
- storage cost, rubles;
- throughput, bit/s and latency, s;
- durability, hours.

The overall system performance can be estimated by indicators such as the average completion time of a typical read or write request (Latency), as well as number of operations per second (IOPS) and number of megabytes per second (MBPS). We assume that for a typical user of a storage, the most understandable and familiar performance indicator from the listed may be the bandwidth or average read or write speed v^{rw} in megabytes per second, since it is the closest to the role of the integrated speed characteristic of a storage system. The specified indicator is defined as the ratio of the total volume of completed requests to the total processing time of the requests.

Of the four main indicators of reliability (reliability, maintainability, durability and storageability), two indicators – reliability and durability – are provided by manufacturers of flash storage devices, which led to the choice of these indicators for evaluating of data storage systems.

The durability of each storage device depends on such an parameter as the TBW (Total Bytes Written), and on the average intensity of the data write or overwrite stream. Disk service time is determined by a device parameter such as Mean Time Between Failures (or $MTTF$, Mean Time To Failure). The reliability and durability of the entire storage system is determined by the number of storage devices and the used RAID level.

In addition, it is of interest to assess such a comprehensive reliability indicator as the availability rate AR, which is defined as the probability of an object to be in working conditions at an arbitrary point in time (except for the planned periods of downtime). This indicator includes a maintainability indicator such as the average system recovery time $MTTR$ (Mean Time to Recovery) and is determined by the ratio $AR = \frac{MTTF}{MTTF+MTTR}$.

The described indicators reflect all the components of effectiveness [8]: productivity, efficiency, and resource usage. Productivity is ensured by a given reliability, capacity and lifetime of the system, efficiency is achieved by a given read and write performance, and resource consumption is provided by the cost of storage. Obviously, from the perspective of improving all the integral indicators of the system, these indicators are mutually exclusive. For more details on conceptual model, see [15].

B. Performance model

The performance model is based on a estimation of the delay time through all the underlying subsystems: the requests are received at the entrance of the storage system by the network service model, at the second stage request goes to the model of the VDO deduplication module, then to the software RAID model and, finally, to the flash memory device model for reading or writing.

A model of user applications create a load on storage. A workload is created by several programs, each of which can request reading or writing. To simplify the model, we assume that the load is from several user programs transmitted over the network to the storage system, is perceived by the network service module at the input of the storage system as a stream of requests.

Each request ($q_i = (t_i, o_i, s_i)$ in the stream is characterized by the following parameters:

- $t_i \in Z$ – time moment of receiving a request (timestamp);
- $o_i \in \{r, w\}$ – requested operation (r - read, w - write);
- $s_i \in N$ – number of bytes (request size) to be processed (written or read).

Main characteristics of the query stream $\{q_i\}$ are:

- $i = 1, \ldots, I$ – number of a request;

- $I \in N$ – total number of requests;
- $t_i,\ i \geq 1$ – time moment of receiving request i by the storage;
- $o_i,\ i \geq 1$ – requested operation of request i;
- $\tau_i = t_{i+1} - t_i,\ i \geq 1,\ t_0 = 0$ – random variables, independent equally distributed time intervals between requests;
- $s_i,\ i \geq 1$ – random variables, independent equally distributed values of requests size;
- t_0 – start time of an experiment;
- t_I – end time of an experiment.

According to the conceptual model presented in II-A, the average execution time of a typical request q_i Δt_i^{rw} is the sum of the service times in each of the components of the system:

$$\Delta t_i^{rw} = \Delta t_i^{NS} + \Delta t_i^{DED} + \Delta t_i^{RAID} + \Delta t_i^{DISK}. \tag{1}$$

Since read and write speed of a flash disk vary significantly, it is reasonable to evaluate two performance indicators: average read speed v^r and average write speed v^w.

Then the total performance indicators of a storage can be written in the following forms:

$$v^r = \frac{\sum_i s_i|(o_i = r)}{\sum_i \Delta t_i^r},\ v^w = \frac{\sum_i s_i|(o_i = w)}{\sum_i \Delta t_i^w}. \tag{2}$$

The delay time at each stage depends both on the characteristics of the requests listed above and on the configuration of the storage system (software and hardware):

- number of CPUs available for request processing at various levels (network service, deduplication),
- amount of RAM available for caching at various levels,
- storage device queuing depth and so called "disk policy" – algorithm of a device queue processing.

We use stochastic simulation to obtain the delay time at each stage of the service request lifetime. Simulation utilities were implemented using R and Python programming language, and SimPy simulation frameworkc̃itesimpy.

The output of the simulation results is implemented by analogy with the output of the `fio` utility, which is widely used to test performance of Linux I/O subsystem. The specified utility allows to get standard integral performance indicators (throughput and IOPS for read or write operations), and also a histogram of the service time distribution for I/O operations.

This is the current work, some draft results are presented in [17].

C. Reliability model

In [9], the first analytical model is proposed for the estimation of reliability dynamics of RAID arrays built on flash storage devices. The authors study the problem of wear-out of a flash disk due to bit errors. One of the main results is the frequency of errors, which depends on time and increases with the level of wear-out of a disk. The efficiency of RAID based on flash disks remains controversial, since parity updates are increasing the wear-out and bit error rate of devices. In the mentioned paper, reliability dynamics of a flash-based RAID is constructed as the solution of a continuous time Markov chain (CTMC) model. The authors take into account various parameters and consider Diff-RAID and RAID 5 as examples.

Unfortunately, this article does not provide results for RAID 6 and RAID 10. In addition, it takes into account only the increasing rate of sudden failures due to bit errors, but it does not take into account the probability of RAID controller or software failure and the probability of disk read errors during data regeneration on replaced disks, as well as the average system recovery time ($MTTR$).

An important consequence of the [9] work is that the failure rate within one operation (read or write) to a flash disk is constant, and the arrival of errors can be modeled by a Poisson process.

Recently, a large number of publications have been devoted to modeling the reliability of a RAID using Markov chains [10], [11], [12]. However, the most of these work does not directly take into account the problem of wear-out.

The basic ideas of the current reliability models for RAID 0, RAID 5, and RAID 6 are first described in [13], while [14] of the same authors presented reliability models for RAID 10 and RAID 01. Mathematical models are Kolmogorov-Chapman systems of equations for calculating stationary probabilities describing transitions between states in a discrete Markov chain.

We modified these models taking into account specifics of a flash disk, such as wear-out that increases the probability of device failure. A model with sequential regeneration of information on replaced disks was considered in a similar way as in [9].

The input parameters of reliability models are following:

- λ – the failure rate of disks in a RAID (the same for all disks);
- μ – data regeneration intensity for a disk in a RAID;
- ε – error rate of a disk read URE in a RAID;
- σ – error rate of a hardware platform and software RAID implementation;
- γ – the intensity of the full recovery of the system from an emergency state;

.

The output parameters of the models are:

- T_F – mean time between failures $MTTF$;
- K_A – availability factor AR;
- T_R – average recovery time $MTTR$.

.

To take flash memory wear-out into account, authors proposed to add the rate of gradual failures due to wearing out λ_W to the rate of sudden disk failures λ_S:

$$\lambda = \lambda_S + \lambda_W \tag{3}$$

.

This assumption is based on the nature of sudden and gradual failures of a flash disk. Device may fail due to both sudden (functional) and gradual failure (result of wearing out). In this case, the total probability of failure is equal to the probability of failure either due to wear-out or a random malfunction, minus the probability of the simultaneous occurrence of these failures.

Assuming a sudden and gradual failure by events joint and independent, the total probability of failure of the flash memory device PF_Σ has the form

$$PF_\Sigma = PF_S + PF_W - PF_S PF_W, \quad (4)$$

where PF_S and PF_W are the probabilities of a sudden and gradual failures of a flash disk respectively.

To calculate the output reliability indicators, one have to evaluate six initial reliability parameters: $\lambda_S, \lambda_W, \mu, \varepsilon, \sigma, \gamma$.

The rate of sudden failures λ_S is estimated based on the hypothesis of the exponential distribution of failures:

$$PF(t) = 1 - e^{\frac{-t}{MTTF}}, t > 0 \quad (5)$$

In this case, the failure rate is inversely proportional to the parameter $MTTF_{DISK}$ – the average time between failures of any of RAID devices:

$$\lambda_S = \frac{1}{MTTF_{DISK}} \quad (6)$$

To determine parameter λ_W, it is necessary to calculate the time of the total device wear-out:

$$T_{TBW} = \frac{TBW_{DISK}(1-\alpha)}{3600 \cdot \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\sum_{k=1}^{i} s_k|(o_k=w)}{E\cdot(t_i-t_0)}\right)} = T_{DSS}^{W}. \quad (7)$$

where TBW_{DISK} is a flash disk endurance metric (“Total Bytes Written”), α – safety factor, E - number of “effective” flash disks in RAID (depends on RAID algorithm), $\sum_{k=1}^{i} s_k|(o_k = w)$ – total amount of bytes written.

Suppose that the working time of a flash disk is limited by the criterion of gradual failures and has an exponential distribution. Then, from the moment the experiment begins, when the system time reaches the time value T_{TBW}, the probability of failure according to the criterion of gradual failures should become close to 1, and the probability of failure-free operation, on the contrary, should be near zero.

Using well-known formula for the exponential distribution:

$$P(T_{TBW} < t_i < \infty) = e^{-\lambda_W \cdot T_{TBW}} - e^{-\lambda_W \cdot \infty} = R_{CR}^{GF} \approx 0, \quad (8)$$

where R_{CR}^{GF} is the critical value of the probability of failure-free operation with which the storage system will function for at least T_{TBW} amount of time.

After applying logarithm:

$$-\lambda_W \cdot T_{TBW} = ln\left(R_{CR}^{GF}\right) \quad (9)$$

Therefore,

$$\lambda_W = \frac{-ln\left(R_{CR}^{GF}\right)}{T_{TBW}} \quad (10)$$

Other parameters are given by a device manufacturer (uncorrectable read error rate ε, hardware platform and software RAID implementation failures σ), or can be estimated using some expert knowledge (recovering from backup intensity γ, and data regeneration intensity after a failed device replacement μ). The more detailed description of the reliability model is given in [16].

D. Cost model

The task of estimating the cost is one of the essential points necessary to design a data storage system.

Let us define the parts that will be included in C_i^{DSS} – the total cost of a storage system ownership:

- C_i^{RES} – cost of resources, $i = 1, \ldots, I$;
- $C_i^{H\&S}$ – cost of a storage system hardware and software, $i = 1, \ldots, I$;
- C_i^{SRV} – cost of a storage maintenance, $i = 1, \ldots, I$;

(C_i^{RES} is calculated by the formula

$$C_i^{RES} = C^{INT} \cdot t_i + C^{ELC} \cdot t_i + C^{RNT} \cdot t_i \quad | \; t_i \in [t_0, t_I], \quad (11)$$

where

- C^{INT} - payment for Internet, [rub / s];
- C^{ELC} - payment for electricity, [rub / s];
- C^{RNT} - rent of premises, [rub / s].

The electricity charge (C^{ELC} depends on the consumption rate C_h^{kW}, [rub / kWh], the power consumption of a server platform and each flash disk, as well as consumption of an air conditioning system:

$$C_i^{ELC} = \frac{C_h^{kW} \cdot t_i}{1000 \cdot 3600}\left(W^{CMP} + nW^{SSD} + W^{CLM}\right) \quad t_i \in [t_0, t_I], \quad (12)$$

where

- W^{CMP} is power consumption of a server platform, [W];
- W^{SSD} is power consumption of a single flash disk, [W];
- W^{CLM} is power consumption of a climate control system [W].

The cost of a data storage system ($C_i^{H\&S}$ includes the cost of hardware components C^{HRD} [rubles] and software on security C^{SFT} [rubles] taking into account the costs of regular updates of licensed software C^{LIC} [rubles / s], as well as costs C^{DES} [rubles] for design and software development:

$$C_i^{H\&S} = C^{HRD} + C^{SFT} + C^{DES} + C^{LIC} \cdot t_i \quad | \; t_i \in [t_0, t_I]. \quad (13)$$

Maintenance costs are estimated by the expression

$$C_i^{SRV} = C^{ADM} \cdot t_i \quad | \; t_i \in [t_0, t_I], \quad (14)$$

where C^{ADM} is the cost of setup and support (rubles / s).

As the result, the total cost of a storage ownership C_i^{DSS} [rub], taking into account the above components, is:

$$C_i^{DSS} = C_i^{RES} + C_i^{H\&S} + C_i^{SRV} \; | \; t_i \in [t_0, t_I], i = 1, \ldots, I. \quad (15)$$

Reliability and cost models were implemented using Python programming language.

III. Simulation results

In this section the simulation results are presented. We used a special experimental all-flash storage to perform validation of the mathematical modesl.

A. Cost model

The largest contribution to the cost is made by a server platform (chassis, motherboard, processors, memory modules, network cards, etc.) and disks.

For a computational experiment, the cost of a storage built on the AIC HA202-PV server platform was taken. AIC HA202-PV which is two server computers in one chassis with common power supplies and access to U.2 form factor disks of NVMe interface. The chassis allows to install a maximum of 24 drives. Network connection of each server computer node provided by 40G QSFP network interface card. Each computer has two Intel Xeon Silver 4110 processors and 64GB of RAM.

Assume that the estimated cost of storage in the configuration described above is 1.2 million rubles. The cost of one flash disk with a capacity of 960 GB is 15000 rubles. Note that similar calculations for the cost of one drive with 10000 rubles and 20000 rubles showed approximately the same results.

The cost of maintenance, including payment for electricity and rent, premises, employees, etc. was not taken into account due to the fact that the value of these costs can vary greatly depending on many factors that are not related to the specifics of a storage area.

The cost of storage by RAID type per terabyte is presented in table I

B. RAID performance

An experiment is conducted to study effect of the number of disks on RAID performance. For this, for three RAID types (RAID 5, RAID 6, RAID 10), the number of devices in the RAID is varied from 4 to 24 (recall, that available chassis supports 24 devices maximum).

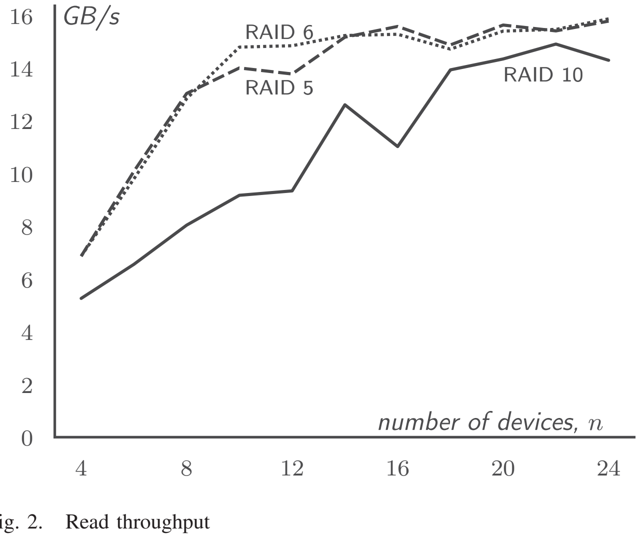


Fig. 2. Read throughput

Fig. 2 and 3 show dependencies between throughput of a RAID and the number of devices in RAID.

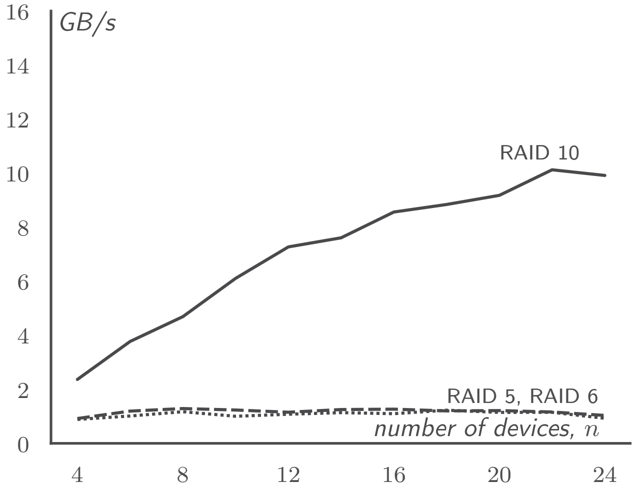


Fig. 3. Write throughput

The read performance for all types of RAID grows with grow in the number of devices in an array; data transfer speed in read mode for RAID 5 and RAID 6 is almost the same.

Read data transfer speed of RAID 10 is lower than that of RAID 5 and RAID 6. However, even RAID 10 reaches the network bandwidth hardware limit (40 Gbit/s or 5 GB/s) with a minimum number of drives.

The behavior of various types of RAID during write workload is significantly different. Due to the nature of their design, RAID 5 and RAID 6 actually have write throughput that is equal to throughput of a single device (about 1 GB/s) with no dependence on a number of underlying disks. With RAID 10, write throughput increases and reaches the hardware limit of the network subsystem for throughput (5GB/s) with ten drives.

C. Reliability model

The reliability model was simulated with the following input parameters:

- $MTTF_{DISK} = 1.8 \cdot 10^6$ hours;
- $P_{UER} = 1 \cdot 10^{-16}$;
- $MTTR_5 = 31/60$ hours (RAID 5 recovery time about 31 minutes);
- $MTTR_6 = 55/60$ hours (RAID 6 recovery time about 55 minutes);
- $MTTR_{10} = 9/60$ hours (RAID 10 recovery time about 9 minutes);
- $MTTR_{DSS} = 24$ hours (empirically, a day for deployment from the backup);
- $MTTE_{CON} = 137592$ hours (from public sources, MTBF for Supermicro SYS-1028U-TR4+ at 20 °C.

Recovery time $MTRR$ after a disk failure for different RAID levels is obtained experimentally for existing equipment (1TB NVMe drives, server on AIC HA202-PV platform, Intel Xeon Silver 4110 processor).

Without taking wear-out into account, reliability is mainly determined by reliability of the server platform (motherboard,

TABLE I. COST OF A DATA STORAGE PER TERABYTE BY RAID TYPE

Number of devices in RAID	Storage devices cost, thousands of rubles	Total cost, thousands of rubles	Size RAID 5 GB	Size RAID 6 GB	Size RAID 10 GB	Cost per TB RAID 5, thousands of rubles/TB	Cost per TB RAID 6, thousands of rubles/TB	Cost per TB RAID 10, thousands of rubles/TB
1	2	3	4	5	6	7	8	9
4	60	1260	2881	1920	1920	437	656	656
6	90	1290	4801	3841	2881	269	336	448
8	120	1320	6721	5761	3841	196	229	344
10	150	1350	8642	7682	4801	156	176	281
12	180	1380	10562	9602	5761	131	144	240
14	210	1410	12483	11522	6721	113	122	210
16	240	1440	14403	13443	7682	100	107	187
18	270	1470	16323	15363	8642	90	96	170
20	300	1500	18244	17284	9602	82	87	156
22	330	1530	20164	19204	10562	76	80	145
24	360	1560	22085	21124	11522	71	74	135

processor, memory, power supply) on which a storage is running. The reliability impact of individual drives is negligible.

With a fixed value of MTBF of a single drive, reliability of RAID 5 is the most depends on the number of disks in the array (the more devices, the less time between failures). For RAID 6 and RAID 10, the effect of the number of devices in the array is significantly less than for RAID 5.

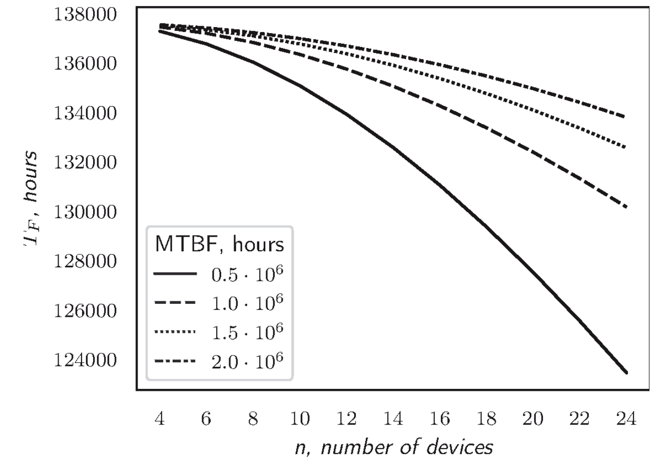


Fig. 4. Mean time to failure T_F for RAID 5 and various MTBF values for underlying disks (without taking into account wear-out)

Most significantly, reliability of a single drive affects the reliability of RAID 5. For this type of RAID, decreasing the reliability of a single drive by 4 times (from $2 \cdot 10^6$ to $0.5 \cdot 10^6$ hours) decreases the time between failures with the maximum number of disks (n = 24) by 7.7%. For the minimum number of disks (n = 4) this difference is only 0.2% (see Fig. 4).

For other types of RAID (RAID 6, RAID 10), reliability impact of a single disk is even less. Decreasing of mean time between failures with the maximum number of disks for RAID 6 and RAID 10 is 0.15% and 0.38% (Fig. 5 and 6).

Fig. 7 shows the same simulation results as Fig. 4 - 6, but using single scale for ease of comparison.

Next, we take MTBF of a separate flash disk of $1.8 \cdot 10^6$ hours.

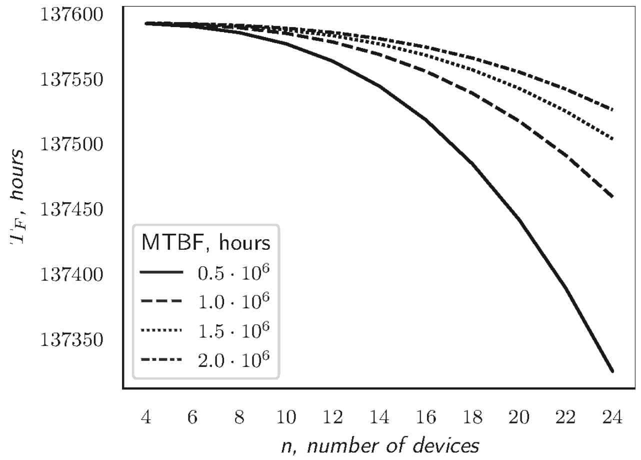


Fig. 5. Mean time to failure T_F for RAID 6 and various MTBF values for underlying disks (without taking into account wear-out)

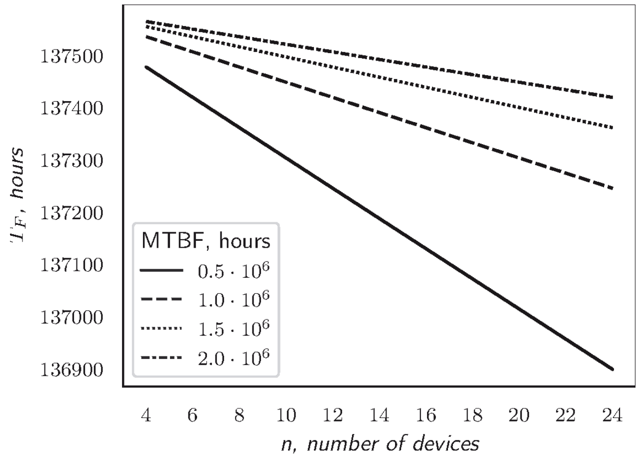
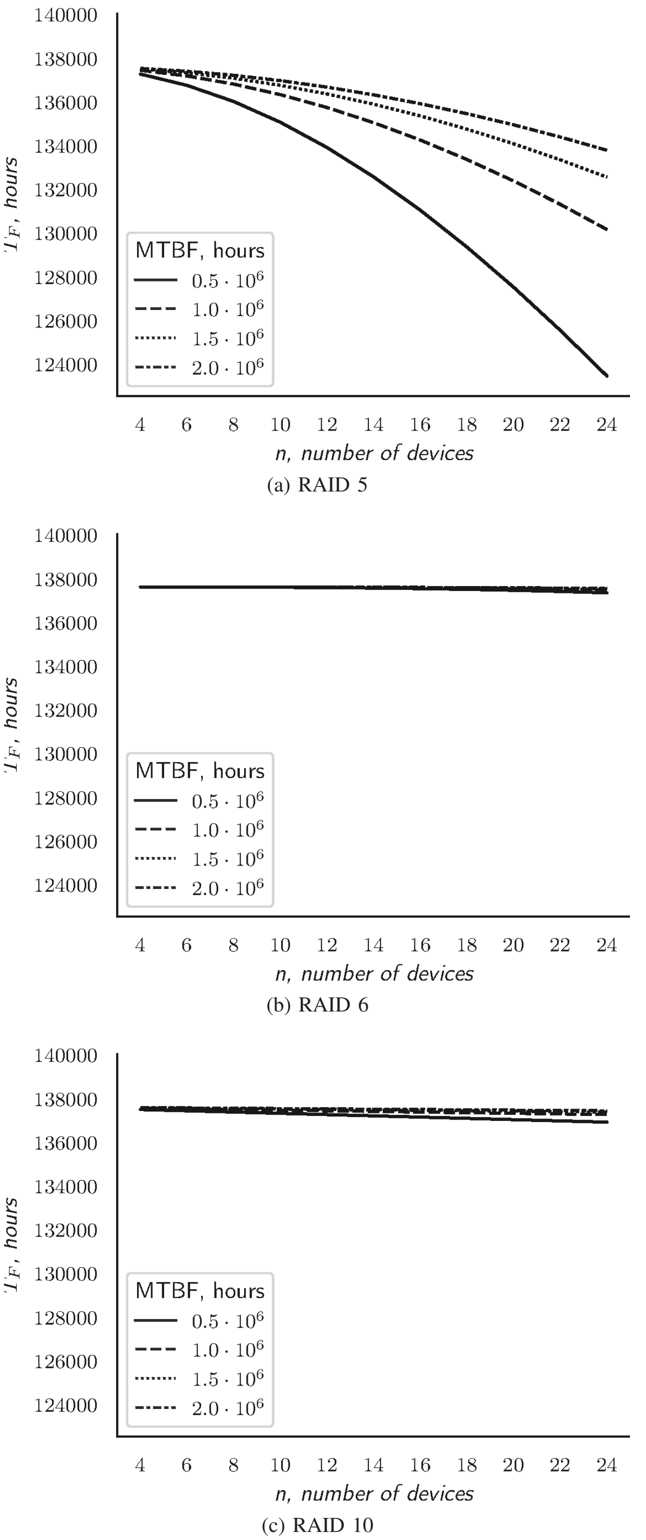


Fig. 6. Mean time to failure T_F for RAID 10 and various MTBF values for underlying disks (without taking into account wear-out)

Fig. 8 shows effect of wear-out for medium intensity (1 Gbit/s) write stream.

When taking into account the effect of wear-out, the mean time before failure (T_F) is reduced:



(a) RAID 5

(b) RAID 6

(c) RAID 10

Fig. 7. Mean time to failure T_F for various MTBF values for underlying disks (without taking into account wear-out)

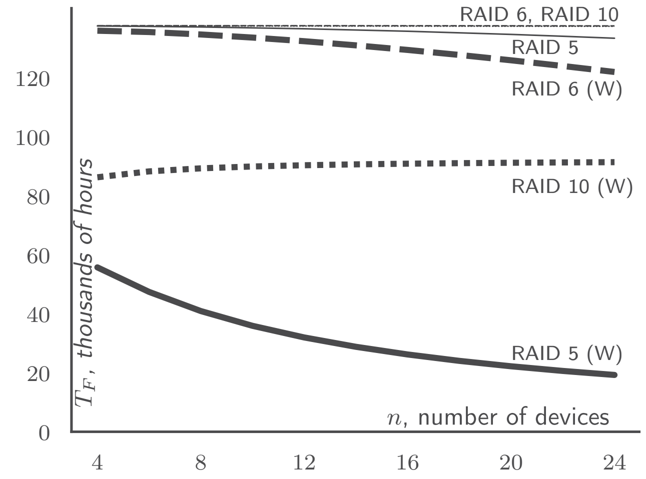


Fig. 8. Mean time to failure T_F for MTBF value $1,8 \cdot 10^6$ hours (taking into account wear-out)

- RAID 6: for an array of four drives from 137592 hours without taking wear-out into account up to 135886 hours, taking wear-out into account (by 1.24%), for 24 devices from 137518 to 121912 (by 1.35%);
- RAID 10: for an array of four drives from 137560 hours without taking wear-out into account to 86225 hours, taking wear-out into account (by 37.32%), for 24 devices from 137399 to 91352 (by 33.51%);
- RAID 5: for an array of four drives from 137495 without taking wear-out into account up to 55706 hours, taking wear-out into account (by 59.5%), for 24 devices to 19251 (by 85.6%).

Thus, the effect of wear is significant. Next figures are only given taking wear-out into account.

Consider the behavior of various types of RAID when the wear-out intensity changes. Wear-out intensity is determined by the average write speed Mv^w. The average write speed, in turn, is determined by the nature of the workload. For comparison, three values of Mv^w were used: 100 Mbit/s (for example, when users copy files over the Internet), 1 Gbit/s (for example, a video surveillance system with a large number of high-resolution cameras), 10 Gbit/s (copy large amounts of data in the data center).

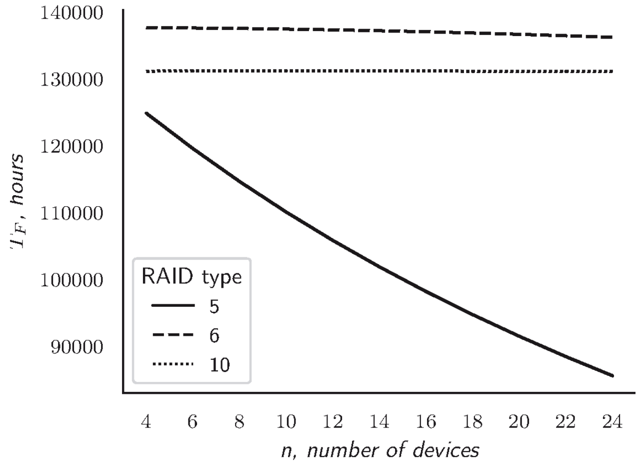


Fig. 9. Mean time to failure T_F taking wear-out into account for write speed Mv^w = 0.1 Gbit/s

The Fig. 9 shows the behavior of various types of RAID at a low level of wear-out (average write speed of 100 Mbit/s).

With a minimum number of disks (n = 4), the mean time between failures is from 137552 hours for RAID 6 to 124757 hours for RAID 5. When the number of disks grows to the maximum (n = 24), the mean time between failures for RAID 6 and RAID 10 vary slightly (1% and less than 1%, respectively). At the same time, the average MTBF of RAID 5 is dropped by noticeable 36%.

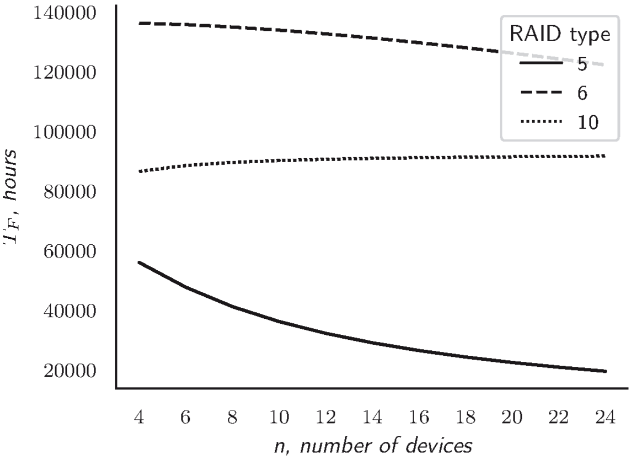


Fig. 10. Mean time to failure T_F taking wear-out into account for write speed Mv^w = 1 Gbit/s

The Fig. 10 shows the behavior of various types of RAID at an average wear-out intensity (average write speed of 1 Gbit/s). With a minimum number of drives (n = 4) MTBF ranges from 135886 hours for RAID 6 to 55706 hours for RAID 5. With the increase in the number of disks to the maximum (n = 24), the mean time between failures RAID 6 decreases by 10%, while RAID 10 increases by 6%. The MTBF of RAID 5 is dropped by 66%.

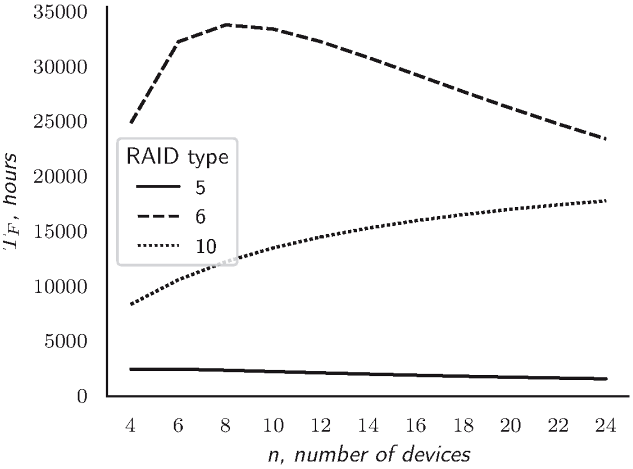


Fig. 11. Mean time to failure T_F taking wear-out into account for write speed Mv^w = 10 Gbit/s

Fig. 11 shows the behavior of different types of RAID at high wear-out intensity (average write speed 10 Gbit/s). With a minimum number of drives (n = 4) MTBF is between 24732 hours for RAID 6 to 2322 hours for RAID 5. With the increase in the number of drives, the mean time between failures of RAID 6 increases and reaches 33667 hours with six drives, after which it decreases to 23284 hours with 24 drives. The average time between failures of RAID 10 with an increase in the number of disks increases almost twice (from 8230 to 17648 hours). For RAID 5, the mean time between failures decreases from 2322 to 1450 hours, which is 16 times less than the time of RAID 6 with the same wear rate of Mv^w = 10 Gbit/s and a comparable size.

Modeling showed that for 24 disks the MTBF taking into account wear-out RAID 6 is 33% longer than for RAID 10. RAID 5 significantly behind RAID 6 and RAID 10.

D. Cost and reliability

Fig. 12 illustrates the results of combining two models: cost and performance. Lines marked “C” shows total cost of storage per Terabyte per year, where storage cost is the result of cost model, and maximum storage time is the result of reliability model (lines marked “D”).

Obviously, the cost of storage per year significantly depends on the nature of the workload, storage size and RAID type. When using a small number of disks (4-5), the unit cost is significant. When storage size increases from 1 to 5 TB, cost per Tb per year decreases by more than 2 times. With an increase in the wear rate to the maximum (average recording speed of 10 Gbit/s), the specific storage cost increases sharply. Lowest cost can be achieved using RAID 6.

IV. Conclusion

In this paper the mathematical models of performance, reliability and cost are presented. We provide the simulation results using these models. The simulation results show that the best value of cost and reliability are obtained using RAID 6; the worst – using RAID 5; RAID 10 is somewhere in intermediate position.

The maximum possible size for RAID 10 is half the maximum capacity for RAID 5 and RAID 6. It is also seen that from the point of view of the cost of storage per TB is inappropriate but use storage to store small amounts of data. Write speed higher that a single disk write speed is possible only with RAID 10.

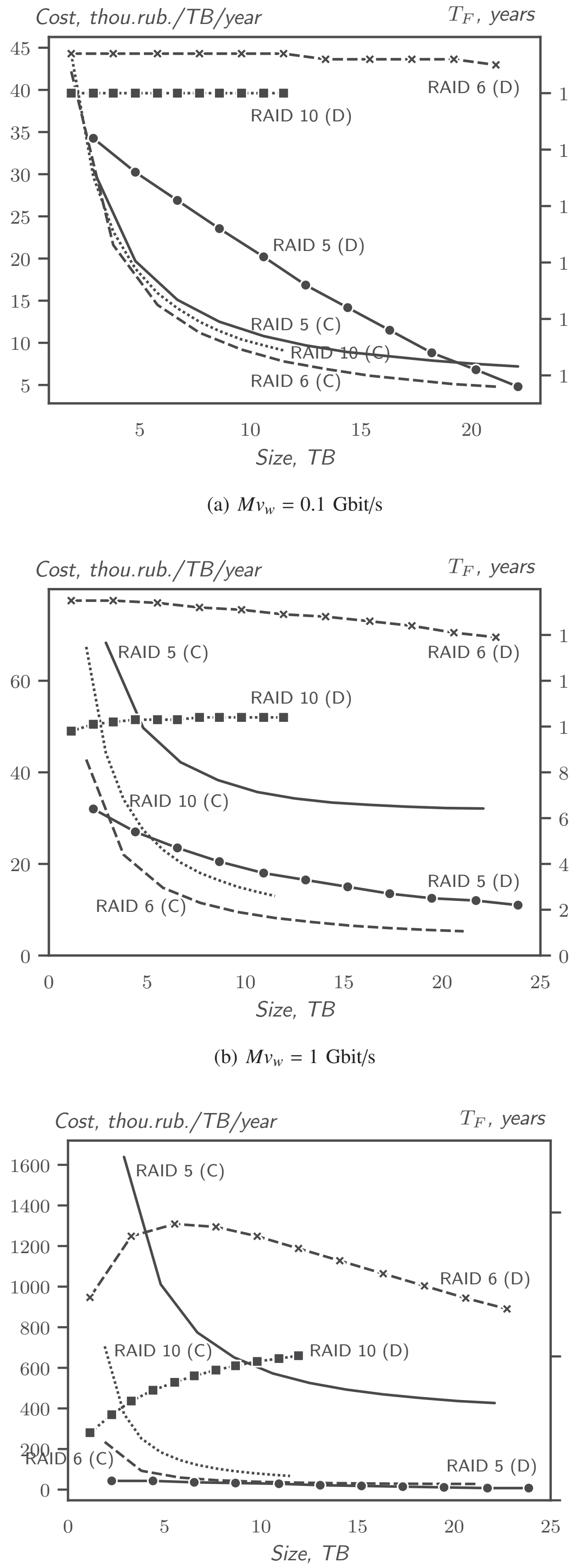
Our further work is improvement of the models to apply for development of smart sensor systems in Internet of Things environments. All-flash data storage is used to maintain sensed data locally, which come from multiple sources. Effective strategies for operation multi-source sensed data are required.

Acknowledgment

This research is financially supported by the Ministry of Science and Higher Education of Russia within project no. 075-11-2019-088 (Creating the high-tech production of mobile microprocessor computing modules based on SiP and PoP technology for smart data collection, mining, and interaction with surrounding sources).

References

[1] The Digitization of the World: From Edge to Core, Web: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf



(a) $Mv_w = 0.1$ Gbit/s

(b) $Mv_w = 1$ Gbit/s

(c) $Mv_w = 10$ Gbit/s

Fig. 12. Cost per TB per year and time to failure for various RAID types

[2] Real world SSD wearout, Web: https://blog.okmeter.io/real-world-ssd-wearout-a3396a35c663

[3] Chernov, Ilya A., Evgeny Ivashko, Dmitry Kositsyn, Vadim Ponomarev, Alexander Rumyantsev and Anton Shabaev, “Flash-Based Storage Deduplication Techniques: A Survey.”, *International Journal of Embedded and Real-Time Communication Systems (IJERTCS)*, 2019, 10.3, pp. 32–48, doi:10.4018/IJERTCS.2019070103

[4] Zhang B., Wang C., Zhou B. B., Yuan D. and Zomaya A. Y., “DCDedupe: Selective deduplication and delta compression with effective routing for distributed storage”, *Journal of Grid Computing*, 2018, 16(2), pp. 195–209, doi:10.1007/s10723-018-9429-3

[5] Data deduplication and compression with vdo, Web: https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/storage_administration_guide/vdo-integration

[6] D. Korzun, E. Balandina, A. Kashevnik, S. Balandin and F. Viola, *Ambient Intelligence Services in IoT Environments: Emerging Research and Opportunities*, Hershey, PA: IGI Global, 2019, doi:10.4018/978-1-5225-8973-0

[7] SimPy: Discrete event simulation for Python, Web: https://simpy.readthedocs.io/

[8] V.S. Anfilatov, A.A. Emelyanov and A.A. Kukushkin, *System analysis in management*. Moscow: Finance and Statistics Publishers, 2002.

[9] Yongkun Li, Patrick P. C. Lee, John C. S. Lui, “Stochastic Analysis on RAID Reliability for Solid-State Drives”, *in Proc. of the IEEE 32nd Symposium on Reliable Distributed Systems*, 2013, pp. 71–80.

[10] P. A. Rahman, G. D’K. Novikova Freyre Shavier, “Analysis of mean time to data loss of fault-tolerant disk arrays RAID-6 based on specialized Markov chain”, *11th IOP Conference Series: Materials Science and Engineering*, vol.327, issue 2, 2018

[11] P. A. Rahman, G. D’K. Novikova Freyre Shavier, “Reliability model of disk arrays RAID-5 with data striping”, *11th IOP Conference Series: Materials Science and Engineering*, vol.327, issue 2, 2018

[12] P. A. Rahman “Using a specialized Markov chain in the reliability model of disk arrays RAID-10 with data mirroring and striping”, *10th IOP Conference Series: Materials Science and Engineering*, vol.177, 2017

[13] P. A. Rahman, A. I. Kayashev, M. I. Sharipov, “Reliability model of fault-tolerant data storage systems”, *Bulletin of the Ufa State Aviation Technical University*, vol.19, №1(67), 2015, pp. 155–166 (in Russian)

[14] P. A. Rahman, E. A. Muraviova, “Markov reliability models of cascade disk arrays RAID-01 and RAID-10”, *Bulletin of a USTU young scientist*, №1, 2015, pp. 52–60 (in Russian)

[15] V. A. Ponomarev, E. A. Pitukhin, “A conceptual model of the functioning of a storage system based on solid-state drives with deduplication technology”, *Engineering Bulletin of Don*, №5, 2019, Web: http://ivdon.ru/ru/magazine/archive/n5y2019/5905 (in Russian)

[16] V. A. Ponomarev, “Mathematical models of performance, reliability and cost of operation of a system for storing deduplicated data on an SSD”, *Engineering Bulletin of Don*, №6, 2019, Web: http://ivdon.ru/ru/magazine/archive/N6y2019/6012 (in Russian)

[17] V. A. Ponomarev, “Simulation Modeling Performance Indicators for Solid State Storage Systems”, *Programmnaya Ingeneria*, 2019, vol.10, №9–10, pp. 367–376. (in Russian)