Determination of parameters of the Automated backup process of digital data for printing houses and publishing houses without use of external network technologies transformations

N E Proskuriakov, B S Yakovlev
Tula State University, Lenin Avenue, 92, Tula, Russia

Abstract. Results of pilot studies of influence of the average size of files, total amount of and volume of the recovered data for the period of backup when using of the automated processes of creation of backup copies of figures for the medium-sized and small printing and publishing enterprises are provided. Experimental and analytical methods revealed the critical parameters influencing the speed of process of backup on condition of use of exclusively internal opportunities of the enterprises. As showed research, the most significant parameters are the optimum size of the file and total amount of data. The option of use of an archiving before direct copying of data is analyzed, positive and negative sides of this approach are shown. The optimum size of the file for process of backup at its use on one computer without use of network technologies and external hard drives was established, its mean value makes 685 Mb. These researches allow increasing quality of backup process; will help creation of the automated, autonomous and flexible systems of backup.

1. Introduction
In modern conditions of printing and packaging products, everything rises a question of data preservation of publishing houses, printing houses, design firms and other establishments more sharply. An increasing number of people reflect on preservation of the practices, ready projects as important components of memory and history of the development. Generally, such tendency appeared because of emergence of the viruses, which capable to destroy information irrevocably, and fragility of data carriers.

Application of classical tools for backup copies which most often for increase in reliability of saving of information include possibility of data storage in the network storages created by large suppliers of similar services would be the most qualitative solution of this problem. However, in the conditions of fierce competition, and at the general adverse background of virus threats, even more often the printing and publishing organizations disconnect this option or are limited to software solutions, trying to be protected from theft of data. For the same reason try not to use external rigid data carriers and not always function of control of versions of files, which is, best of all implemented in network or «cloudy» storages works.

To reduce risks of loss of data in similar conditions, long ago and are successfully applied different program and the hardware-automated tools for backup copies. Conditionally, these technologies can be divided into the following types:
1. Ensuring fault tolerance at storage of data;
2. Ensuring reliability of preservation of data;
3. The symbiosis of the solutions described above.
Still both simple users, and a certain percent of system administrators and engineers confuse fault-tolerant system of data storage to concept reliability of storage, meaning that this same. However, it is approaches very different from each other, which solve different problems.

The fault-tolerant system solves a problem of maintenance of operability of the server or any device in case of loss of part of physical carriers on which valuable information contains. It is implemented due to introduction of RAID of arrays the most widespread of which are RAID-1, 5, 10 [1-3]. The basis of RAID-arrays levels 5 and 10 consists available a large number of the hard drives identical according to characteristics, volume, division into the logical disks united in uniform «virtual» space in which information registers afterwards. Thus one of data carriers is used as reserve, stores in itself the list of files, their checksums, and in case of refusal one of hard drives uses reserve, writing down on it data, thereby recovering working capacity or availability of information. RAID-1 works a little differently, it does not create «virtual» space, and two hard drives are full copies of each other.

In case of the computer infection with a virus damaging data similar fault-tolerant systems will save the infected files, there will be no opportunity «to roll back» their versions, thereby the fault-tolerant system does not protect from virus threats at all, it was not created for this purpose.

Reinstallation of an operating system or failure of the hardware RAID-controller since these situations can lead to recreation of RAID of an array and total loss of data will be a problem for saving of data on RAID-arrays [2]. This problem especially sharply rises in the presence of not enough qualified personnel who can make mistakes, fatal to the stored information, owing to the inexperience.

Ensuring reliability of saving of data is very variable technology, approach that has to provide the user with opportunity to leave with the minimum losses or without them in case of infection with viruses or wrong removal of information. In modern realities, usually, such method is carried out due to storage of files on data carriers, different in an arrangement, for example, external and internal hard drives, involves network technologies, for example, cloudy file storages of data - "clouds", local file servers, NAS storages and other technologies.

It should be noted that reliability of storage in similar approach directly depends on that as far as actions for data security, procedures were thought over, and how often there is a creation of backup copies [1]. Recently network storages start entering the list of obligatory measures for prevention of loss of data.

For this purpose, there are three reasons:

1. Impossibility to damage this service physically because of its remoteness;
2. The impossibility of distribution of a virus in storage since it works at UNIX to a similar operating system and does not give to the user opportunity to establish or start something in storage;
3. Availability of opportunity «to be rolled back» to earlier version of the file, even in case of full synchronization after infection.

In our opinion, it is necessary to aim at preservation of these technologies given by symbiosis as it sharply increases chances of recovery of data. However, to enter fault-tolerant systems everywhere it is expensive and, in most cases unfairly. Arrays it makes sense to enter RAID only on the separate computer, which acts in a local network as the server, the file server or has other, important application.

Practically all system administrators, known that process of backup is very long, especially, if it is not optimized and not automated. Therefore, we consider that an important task in improvement of quality of creation of backup copies is acceleration of this process, reduction of time of operation.

Earlier researches on the organization of data storage taking into account use of network storages were published [4]. In them the necessary facts and measures that can help with a choice of network storage, with achievement of the greatest speed of process of synchronization were determined. It was established that it is more reasonable to create multivolume archives, using the archiver 7z, with their division into volumes according to 300 Mb.

Objective of this research was definition of influence of different parameters on speed of the automated process of backup when using standard and specialized free software of data synchronization in storage files conditions, using only internal hard drives of one computer.
The hypothesis of the influence of the average file size, the removed volume and the total volume on the time and backup speed was checked.

2. Objects and methods of experimental researches

For definition of influence of the average size of files on the process of speed backup catalogs were used with a capacity of 30 Gb. These catalogs consisted of files sets of the identical size: 100 Kb; 1, 10, 100, 500 Mb and 1 Gb. Files were received by the command: `fsutil.exe filecreatenew team <a file name, including a way to it> <the size in bytes>`.

The catalogs received thus comprised the following number of subdirectories and files (see Table 1)

Table 1. Structure of test catalogs for check of a hypothesis of influence of the average size of files on the speed of backup process of taking into account the total volume of 30 Gb

| Average size of the file | Number of catalogs and subdirectories | Number of files |
|-------------------------|---------------------------------------|-----------------|
| 1 Gb                    | 3                                     | 30              |
| 500 Mb                  | 40                                    | 60              |
| 100 Mb                  | 100                                   | 300             |
| 10 Mb                   | 394                                   | 3 000           |
| 1 Mb                    | 3 394                                 | 30 000          |
| 100 Kb                  | 33 400                                | 300 000         |

Researches were conducted on the computer with the following characteristics:
1. Processor: Intel(R) Core(TM) i7-3770 CPU @ 3.40 GHz
2. Random access memory: 8 Gb
3. Type of system: 64-digit Windows 7 PRO operating system
4. Hard drive WDC WD30EFRX-68EUZN0 SATA/300 of 3 of Tb
5. Hard drive ST3000DM001-1E6166 SATA/300 of 3 of Tb.

As instruments of check to use utilities of Robocopy and cwRsync, which showed good efficiency in former work [4].

Utilities were started in the mode of a command line from files of CMD-format:
1. `robocopy "way to the catalog with samples" "*.*" "a way to the checked catalog" / E/MIR/LOG: a way to the file with the report of .log`;
2. `robocopy "way to the catalog with samples" "*.*" "a way to the checked catalog" / E/LOG: a way to the file with the report of .log`;
3. "A way to the utility (cwRsync\bin\rsync.exe)" - av - delete - update - info=progress2 - log-file = "a way to the file with the report of .log" "a way to the catalog with samples" "a way to the checked catalog".

It is worth explaining that the utility of Robocopy was started in two modes with parameters:
/ E (simple copying, without change of structure of the violated catalog);
/ E/MIR (the mirroring mode when the checked catalog can change depending on original).
It was authorized to utility of cwRsync to delete and update catalogs and files in the checked catalog. Both utilities wrote down reports of operation in the LOG file.

For establishment of influence of the deleted volume of data «loss» of files was emulated, that is removal of a certain volume of data was carried out. There were removed as small data on volume - 100, 500 Kb, and 10, 20, 30 Gb.

Table 2. Time of the analysis and recovery of 500 Mb of data at the total volume of 30 Gb

| Average size of the file | Programs (utilities) |
|-------------------------|----------------------|
| 1 Gb                    | Robocopy with key /E | Robocopy with key /E/MIR | cwRsync |
|                         | –                     | –                         | –       |
Results of experiments are provided in Tables 2–9, where the operating time of each utility is given in minutes.

**Table 3. General time of the analysis and recovery of 1 Gb of data from files with a capacity of 30 Gb**

| Average size of the file | Programs (utilities) | Robocopy with key /E | Robocopy with key /E/MIR | cwRsync |
|--------------------------|----------------------|-----------------------|--------------------------|---------|
| 1 Gb                     | 0.02                 | 0.02                  | 0.18                     |
| 500 Mb                   | 0.08                 | 0.08                  | 0.22                     |
| 100 Mb                   | 0.08                 | 0.08                  | 0.25                     |
| 10 Mb                    | 0.15                 | 0.13                  | 0.25                     |
| 1 Mb                     | 0.20                 | 0.22                  | 0.87                     |
| 100 Kb                   | 3.85                 | 4.02                  | 8.27                     |

**Table 4. General time of the analysis and recovery of 10 Gb of data from files with a capacity of 30 Gb**

| Average size of the file | Programs (utilities) | Robocopy with key /E | Robocopy with key /E/MIR | cwRsync |
|--------------------------|----------------------|-----------------------|--------------------------|---------|
| 1 Gb                     | 1.60                 | 1.62                  | 1.92                     |
| 500 Mb                   | 1.70                 | 1.72                  | 2.02                     |
| 100 Mb                   | 1.62                 | 1.65                  | 2.18                     |
| 10 Mb                    | 1.75                 | 1.75                  | 2.13                     |
| 1 Mb                     | 1.92                 | 1.92                  | 5.97                     |
| 100 Kb                   | 8.57                 | 8.45                  | 42.63                    |

**Table 5. General time of the analysis and recovery of 20 Gb of data from files with a capacity of 30 Gb**

| Average size of the file | Programs (utilities) | Robocopy with key /E | Robocopy with key /E/MIR | cwRsync |
|--------------------------|----------------------|-----------------------|--------------------------|---------|
| 1 Gb                     | 3.43                 | 3.38                  | 3.92                     |
| 500 Mb                   | 3.45                 | 3.47                  | 4.05                     |
| 100 Mb                   | 3.50                 | 3.53                  | 4.40                     |
| 10 Mb                    | 3.47                 | 3.45                  | 4.17                     |
| 1 Mb                     | 3.87                 | 3.97                  | 11.28                    |
| 100 Kb                   | 13.30                | 13.30                 | 86.77                    |

**Table 6. General time of the analysis and recovery of 30 Gb of data (elimination of data)**

| Average size of the file | Programs (utilities) | Robocopy with key /E | Robocopy with key /E/MIR | cwRsync |
|--------------------------|----------------------|-----------------------|--------------------------|---------|
| 1 Gb                     | 5.28                 | 5.23                  | 5.77                     |
| 500 Mb                   | 5.30                 | 5.28                  | 6.28                     |
| 100 Mb                   | 5.32                 | 5.23                  | 6.72                     |
Further additional experiments, which had to define influence of the deleted small volume of data on process speed indicators, were made. Experiments were made with the catalogs prepared of the smallest files on the average overview, modeling the worst option on operation time. These results are given in Table 7.

**Table 7. Influence of quantity and volume of files on a recovery time of data (for 100 Kb)**

| Number of remote files (It is removed in a random way) | Programs (utilities) | Programs (utilities) | Programs (utilities) |
|------------------------------------------------------|----------------------|----------------------|----------------------|
|                                                      | Robocopy with key /E | Robocopy with key /E/MIR | cwRsync |
| One file, 100 Kb                                    | 1.25                 | 1.23                 | 4.55                |
| Two files, 200 Kb                                   | 1.27                 | 2.33                 | 4.52                |
| Three files, 300 Kb                                 | 1.27                 | 2.50                 | 4.52                |
| Four files, 400 Kb                                  | 1.23                 | 2.65                 | 4.53                |
| Five files, 500 Kb                                  | 1.30                 | 2.78                 | 4.62                |

Then ability of utilities to recovery of structure of catalogs was investigated. Modeling of violation of structure consisted in full removal of files from the catalog and, in the second option, addition to the available catalog in 30 Gb of other catalog similar on the total amount (Table 8).

**Table 8. Influence of quantity and volume of files on a recovery time of data (for 100 Kb)**

| Number of files | Programs (utilities) | Programs (utilities) |
|-----------------|----------------------|----------------------|
|                 | Robocopy with key /E/MIR | cwRsync |
| Primary preparation of 30 Gb, (300 000 files, catalog empty) | 18.52 | 128.48 |
| Check on mistakes (30 Gb of wrong information). The total volume of the catalog is 60 Gb (600 000 files) | 10.13 | 13.68 |

To reduce the time of archiving, the hypothesis of application of archiving without compression was checked. Results of experiment are given in Table 9.

**Table 9. The spent time for operation of copying of multivolume archive of the catalog with a capacity of 30 Gb with breakdown of files in volume 500 Mb**

| Type of the utility | The size of files in the catalog | Time of archiving operation | Time of archive synchronization | The total time of the operation |
|---------------------|----------------------------------|-----------------------------|--------------------------------|--------------------------------|
| Robocopy with key /E/MIR | 100 Kб | 11.42 | | 16.70 |
|                      | 1 Мб  | 6.12  | | 11.40 |
|                      | 10 Мб | 6.02  | | 11.30 |
|                      | 100 Мб| 5.97  | 5.28 | 11.25 |
|                      | 500 Мб| 6.0   | | 11.28 |
|                      | 1 Гб  | 5.97  | | 11.25 |
| cwRsync             | 100 Кб | 11.42 | | 17.65 |
|                      | 1 Мб  | 6.12  | | 12.35 |
|                      | 10 Мб | 6.02  | | 12.25 |
|                      | 100 Мб| 5.97  | 6.23 | 12.20 |
|                      | 500 Мб| 6.0   | | 12.23 |
|                      | 1 Гб  | 5.97  | | 12.20 |
3. Statistical processing and approximation of experiments results

Statistical processing of experimental data was carried out with use of the integrated MATLAB package and its standard polyfit program. Based on Table 3-7 the following graphic dependences of indicators of average speeds of backup (see Figure 1-3) were received. The graphic dependence illustrating the speed of process of backup based on results of Table 9 is presented on Figure 4.

![Figure 1. Backup speed indicators for 10 GB data recovery](image1)

![Figure 2. Backup speed indicators for 20 GB data recovery](image2)

![Figure 3. Backup speed indicators for 30 GB data recovery](image3)

![Figure 4. Indicators of copying speed of multivolume archive of the catalog with a capacity of 30 Gb with breakdown in 500 Mb](image4)

4. Results and discussions

The analysis of Tables 2-6 revealed insignificant influence of the average size of files on process speed. All graphic dependences (Figure 1-4) of backup operation speed show availability of a maximum on curves of speeds. For different utilities, these points of a maximum of speed of process have different values. According to the given dependences, the optimum size of the file, which is necessary for achievement of the highest, speed, has the size from 650 to 720 Mb that averages 685 Mb.
In Figure 1-3 steady lag on the speed of the utility of cwRsync from Robocopy irrespective of, with what parameters they were started is shown. Also from graphic dependences, it is visible that the difference on copying speed between different modes of behavior of the utility of Robocopy is insignificant.

Despite that, in general by all results of experiments of cwRsync showed the inefficiency on process speed, it in certain cases is obligatory to application. The last is connected with that the utility of Robocopy does not cope with exceeding of length of the file name (more than 250 symbols) and passes them, without asking and without warning the user [5, 6]. It means that at synchronization of not archived catalogs, files can be lost, that is inadmissible In this case application of cwRsync will be the only option. It is also noticed that cwRsync precisely repeats time and date of creation of files whereas Robocopy determines date of copying of the catalog.

The Table 7 shows that at recovery of data with a capacity from 100 to 400 Kb time of operation is approximately identical and changeable. Only a volume in 500 Kb has insignificant impact on the general process, and reveals a tendency on increase in time of operation.

The Table 8 represents dependence of quantity and volume of files on a recovery time of data at availability or absence of files in the catalog is initial. It is shown that operation happens much quicker if the checked catalog already has files inside.

The analysis of experiment data shows in Table 9 that approach using an archiving is the best of all works in case of availability of large volumes of data, files, consisting from rather small by size. Otherwise because of availability of operation of an archiving the speed of process of backup decreases and there is more slowly, than a synchronization of catalogs in an initial state.

5. Conclusions

The analysis of research results (Table 7 and Figure 1-3) shows that the number of files and volume of the deleted or «lost» files have the greatest influence on process of backup.

The optimum size of the file for process of backup at its use on one computer and without use of network technologies and external hard drives was established, its average value makes 685 Mb.

Introduction to operation of process of an archiving not always positively affects backup process time. Further researches, are necessary for identification of thresholds at which it is more reasonable to enter this operation. However, despite it, the archiving can be applied all the same since it in most cases is necessary for convenience of a file transfer and carries utilitarian assignment, and also creates conditions for rather cheap option of the automated backup process [7, 8].

These researches will allow increasing quality, convenience and stability of operations in the course of backup in the local, isolated conditions and will help developers to create further reliable automated, autonomous and flexible systems of backup.

References

[1] Ruben G A 2011 How to automatically test and validate your database backup and recovery strategy J. Phys.: Conf. Ser. 331 042031 http://dx.doi.org/10.1088/1742-6596/331/4/042031
[2] P A Rahman and G D K Novikova Freyre Shavier 2018 Reliability model of disk arrays RAID-5 with data striping IOP Conf. Ser.: Mater. Sci. Eng. 327 022087. http://dx.doi.org/10.1088/1757-899X/327/2/022087
[3] Rahman P A 2017 Using a specialized Markov chain in the reliability model of disk arrays RAID-10 with data mirroring and striping IOP Conf. Ser.: Mater. Sci. Eng. 177 012087. doi:10.1088/1757-899X/177/1/012087
[4] Proskuriakov N E, Yakovlev B S and V V Pries 2018 Improving the Quality of Backup Process for Publishing Houses and Printing Houses J. Phys.: Conf. Ser. 998 012026. http://dx.doi.org/10.1088/1742-6596/998/1/012026
[5] Mikhail gotch Where my files, ROBOCOPY? [https://habr.com/post/261359/]
[6] Malikov V We use Robocopy. Examples. [https://vladimirmalikov.com/%D0%B8%D1%81%D0%BF%D0%BE%D0%BB%D1%8C%D0%B7%D1%83%D0%B5%D0%BC-robocopy-]
[7] D Mosk Work with 7zip from the command line. https://www.dmosk.ru/mininstruktions.php?mini=7zip-cmd

[8] Pakhomov S Comparison of 64-bit archivers WinRAR 4.2, WinZip 17.0 and 7-Zip 9.30. https://compress.ru/article.aspx?id=23664