Belarusian State Technological University
Department of Information Systems and Technology

Pavel Urbanovich

INFORMATION PROTECTION

Part 3: BACKUP COPY. **RAID** TECHNOLOGY

pav.urb@yandex.by, p.urbanovich@belstu.by

The most effective protection against data destruction is regular creation of backups.

Backup Copy - a copy of the data, allowing to reproduce the original data in the event of their destruction, damage or loss.

- The backup creation mechanism is very widespread.

Types of Backups

1. Full - full copy contains all the data necessary for them correct reproduction.
   Advantage - requires only the last set of copies.
   Disadvantage - large size.

2. Incremental - contains only data created or changed since creation last incremental or full copy.
   Advantage - reducing copying time.
   Disadvantages:

- make a full backup before creating the incremental backup security copy,
- you need to have a full backup and all incremental backups in chronological order to restore them.

3. Differential - contains only data created or changed since the last full backup.

Advantage - reduced copy time.

Disadvantages:

- create a full backup before making a first copy,
- it is necessary to have a (initial, starting) full backup copy and a last copy of the differential copy.

4. Normal - contains copies of files without changing their archiving attributes.

Advantage – allows to create a backup without violation the integrity of incremental backups.

5. Daily - contains a copy of the files modified on a given day, also without changing the attributes of archiving.

Advantage - is useful for backing up a given day, no violation of the integrity of the incremental backups performed.

Backup Carriers

- Backups may be stored on different carriers depending on the requirements and capabilities of the backups:

1. Magnetic Tape - is the most commonly used carrier for archiving wholesale quantities of data.

➢ **Advantages:**
- high Carrier capacity,
- quite low price per 1 GB.

➢ **Disadvantages:**
- special software is required,
- the tape drive may only be connected to one computer.

The latest magnetic tapes are characterized by higher write speeds than modern hard drives.

2. Hard Drive

➢**Advantages:**

- short access time,
- availability,
- big opportunities,
- ease of use.

➢**Disadvantages:**

- high vulnerability to damage especially during transport.

3. Optical Disc - ordinary CD (DVD) can be used as a data carrier.

➢**Advantages:**

- universality (now almost every computer is equipped with a burner optical discs),

CD burner *- a device for producing a compact disc by copying from an original or master copy*

-low price.

➢**Disadvantages:**

- their durability, after a few years, there may be problems with reading data stored on them.

4. USB Memory - devices based on semiconductor structures.

➢**Advantages:**
- small size,
- seamless mobility.

➢**Disadvantage:**
- relatively expensive in terms of amount of space they offer.

5. Remote Storage Carriers – it's possible a disc connected to the network in the model Network Attached Storage (NAS) to use for backup from multiple servers and workstations running on the network.

➢**Advantages:**
- due to the fact that such a disc is always on, this operation can be performed automatically (using special backup software, sometimes included with the device),
- physical isolation of the backup from the original data.

➢**Disadvantages:**
- high network burden during backup creation,
- not the lowest price of network discs.

Access Methods (for Backup Carriers)

1. On-Line - characterized by very short access time (measured in milliseconds).

A typical example is an internal hard disk (HDD) or disk array.

2. Near-Line Media - they are slower than the previous type, but cheaper.

A typical example is a tape library where access time varies from a few seconds to several minutes.

Tape libraries have 3 essential elements:

- storage carriers (tapes or discs),
- carriers drive (s),
- a robot that allows to pick up the carrier from the magazine and load it into the drive and vice versa.

Methods of Copy Data Creation

1. Simple Copying - the simplest and most common method for backing up.

- This type of functionality has all the backup programs and all operating systems.

2. Partial Copy Copying - instead of copying entire files, you can limit only copying blocks or even bytes that have been changed (over a given period of time).

- This technique requires less storage spacefor back up.

- On the other hand, the retrieval data process is more complicated than in the first basic model.

3. Copying a whole File System

- A typical example is the creation of **disk images**.

- This process usually requires exclusive access to storage space.

- This process is faster than simple file copying.

- Disk images - a file containing a complete copy of the contents and structure of the file system and the data on the disk
- a typical backup program saves only files that you can access; the boot loader and files blocked by the operating system may not be saved
- disk image contains all the data available on the disk

4. Identification of Changes – some file systems have an archive bit that specifies whether the file was last changed.

- This information is used to make incremental or differential backups.

5. Versioning File System

- File systems generate and store a history of all file changes, giving the ability to go back to any version of a file.

- An example is the Wayback versioning file system running on Linux.

Processing and Optimization of Backups

1. Compression

– is an optimization mechanism that allows you to reduce the size of backups,

- all major backup systems have this functionality; in addition, hardware devices such as tape drives also have the ability to compress.

2. De-duplication

- many similar backup systems write data on the same device, it comes to the situation where the data is duplicated;
- for example, if several Windows servers back up to the same location, they could share a common part of duplicate data. In this situation, only one data set is stored.

3. Duplication

- it is often necessary to create additional backups to increase security,
- sometimes, during this operation, we can optimize the copy image to speed up a possible recovery operation.

4. Encryption

- backups of removable disks can be lost or stolen,
- an effective way to protect your data from unauthorized access is through encryption,
- unfortunately, cryptographic operations are quite costly to compute slows down the backup process.

5. Multiplexing

- the data carrier should guarantee the ability to create copies on several systems independently.

6. Refactoring

- the process of making changes to backups,
- for example, if your system makes incremental backups on multiple systems on different tapes, you will need to use many different tapes in case of a failure,

- refactoring and programming - is the process of redesigning of a code to improve its readability and performance, as well as to improve security.

7. Staging

- some backup operations write data first on the intermediate disk and then on the target tape, this process is often referred to as D2D2T (acronym Disk to Disk to Tap),
- this operation is used when the problem is with the speed of writing data on the tape.

External Software to Create Backups

In addition to the system backup tool, there are quite a rich collection of external software:

AcronisTrue Image Home 2010
Auto-Backup –(freeware)
Bekaper–(freeware)
Cobian Backup – (freeware)
DriveImage XML – (freeware)
Fbackup – (freeware)
Ferro Backup System
Handy Backup
Idle Time Backup –(freeware)
Lomsel Backup
MyBackup Pro
Norton Ghost
Norton Save & Restore
Peter's Backup –GPL
Power Folder
SyncBackFreeware –(freeware)
WinImage
WinPSF

II. RAID Technology

RAID - Redundant Array of Independent Disks

The Motivation for RAID

- Computing speeds double every 3 years
- Disk speeds can't keep up
- Data needs higher **MTBF** than any component in system **MTBF** - **Mean Time Between Failures**

-It's a technology that enables greater levels of performance, reliability and/or large volumes when dealing with data.

-By **concurrent use** of two or more 'hard disk drives'.

-**Error correction** (based on Redundancy or Parity Redundancy) techniques combined with multiple disk arrays give you the reliability and performance.

- RAID technology includes many schematics for the construction and operation of disk arrays.
- The name of each schema begins with the word RAID, and then counts the specified schema (eg. RAID 0, RAID 1, …).

Commonly used ones:
RAID 0,
RAID 1,
RAID 5,
RAID 10,
Other types used…but rarely: RAID 2,3,4,6,50……

- Several disk drives using RAID technology are called **RAID arrays.**

Parity Redundancy

- **Parity = XOR** of data from every disk in the RAID unit
- **XOR – eXcluding OR**

Let's we use a matrix of three disks (RAID 4).

The first 2 store the following 8-bit strings:

- disk 1: 01101101
- disk 2: 11010100
- In contrast, disk 3 will be used to store **parity data**:

  01101101 **XOR** 11010100 = 10111001

- Let's assume that the disk 2 has been corrupted by , then by performing the XOR operation on data on disks 1 and 3 we get lost data from disk 2:

  10111001 **XOR** 01101101 = 11010100

RAID 0

a. It splits data among <u>two or more disks</u>.

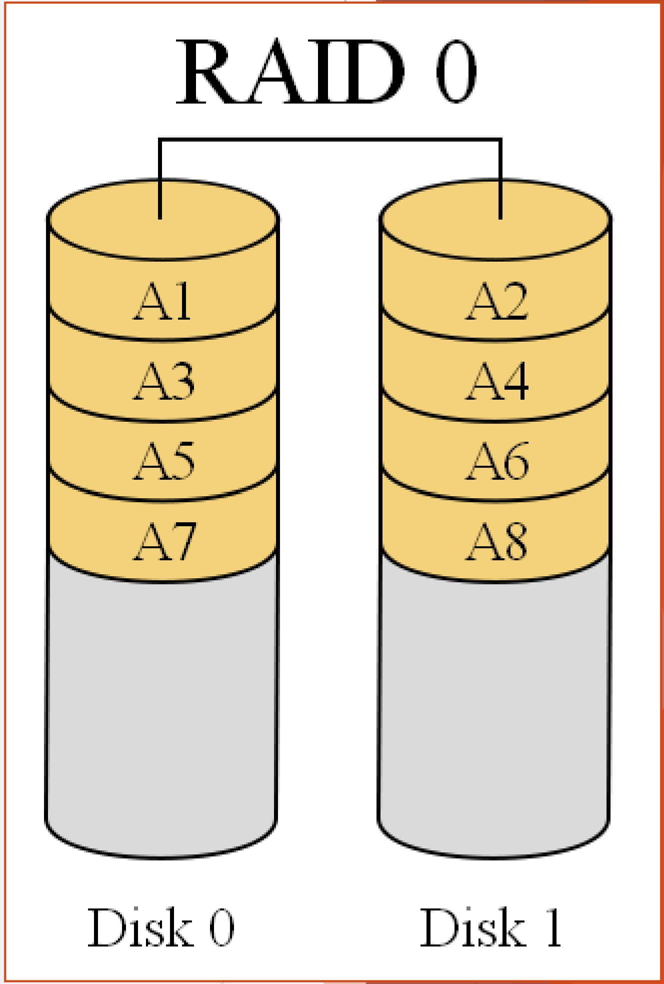
b. Provides <u>good performance</u>.

c. No data redundancy, it means there is no fail over support with this configuration.

d. In the diagram to the right, the odd blocks are written to disk 0 and the even blocks to disk 1 such that A1, A2, A3, A4, … would be the order of blocks read if read sequentially from the beginning.

e. Used in read only NFS systems and gaming systems;

NFS - Network File System



- **RAID 0** distributes data to two or more disks.
So connected physical disks are seen by the user as one disk.

The resulting space has the same size:
size the smallest of the discs * disk quantity.

- **RAID 0** does not provide any redundant information and therefore, we do not get any additional resistance to the errors.

- A failure of one of the disks causes failure of the whole array.

Advantages:
- The space of all discs is seen as a whole,
- Write and read acceleration in comparison to a single disc.

Disadvantages:
- Lack of fault tolerance in the event of a failure of one we lose all data,
- Capacity limited to the **number of discs * size of the smallest discs,**
- Higher probability of failure.

Example

Three 320GB drives have been connected to **RAID 0**.

- The resulting space is 960 GB in size.
- Write or read speed is almost three times as great single disk.
- Total speed is 3 times slowest disk speed.

Example

Three disks: 160 GB, 320 GB and 80 GB were connected in **RAID 0**.

- The resulting space has the same size as the disk size * size the smallest of the disks, that is 3 · 80 GB = 240 GB.
- The speed is limited by the speed of the slowest disk, analogically to the previous example.

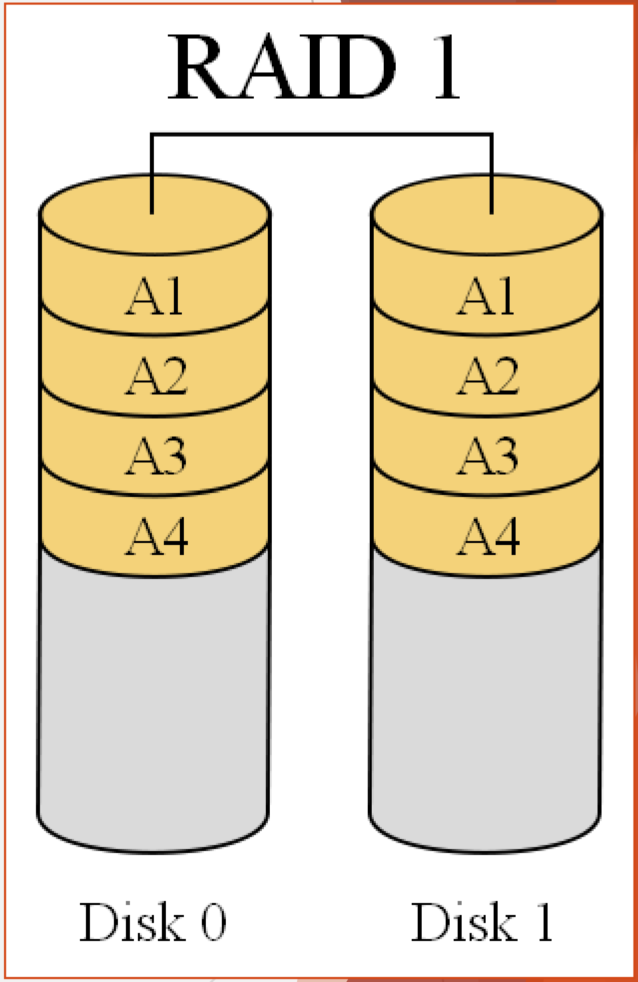
RAID 1

-**RAID 1** is ‘data mirroring’.

- Two copies of the data are held on two physical disks, and the data is always identical.

- Twice as many disks are required to store the same data when compared to **RAID 0**.

- Array continues to operate so long as at least one drive is functioning.



Example

Three 320 GB disks were connected in **RAID 1**.
The resulting space is 320 GB in size.
One or two disks are damaged at some point - the whole array is still working.

RAID 2

- The data on the disks are very finely stripped, even with **1 bit per strip**.

- We need a minimum of 8 data drives and additional disks to store information generated using **Hamming code** needed for error correction.

- Available capacity is the sum of the storage capacities.

Advantages:

- any disk (both with data and Hamming code) in case of damage, it will be rebuilt by other disks.

Disadvantages:

- it is necessary to accurately synchronize all disks containing Hamming code;
  long-term code generation Hamming.

RAID 3

- Data is stored on N-1 disks. The last disk (N-y) is used to store checksums.

- It works like striping in (**RAID 0**), but an additional disk is placed in the array, on which the parity codes are calculated by the special processor.

Advantages:

- fault tolerance 1 disk,
- increased read speed.

Disadvantages:

- in the event of a disk crash, data access is slow because of checksums,
- a single, isolated drive on checksums - usually a bottleneck in the performance of the entire array.

Example

Five 250GB disks were connected in **RAID 3**.

The resulting space has a size of 1TB (250 GB reserved for checksums).

One disk at a time is damaged. The whole array is still working.

When a new disk is inserted into a damaged location, its contents are restored.

RAID 4

- It is very close to **RAID 3**, with the difference that the data is shared on larger blocks (16, 32, 64 or 128 kB).

- Packets are written on disks similar to RAID 0. For each row of data written, the parity block is written to the parity disk.

- Damage to the disc may be reproduced by appropriate mathematical operations.

- **RAID 4** is very good for sequential write and read data (operations on very large files).

- One-time write of small data requires modification appropriate parity blocks for each I/O operation.

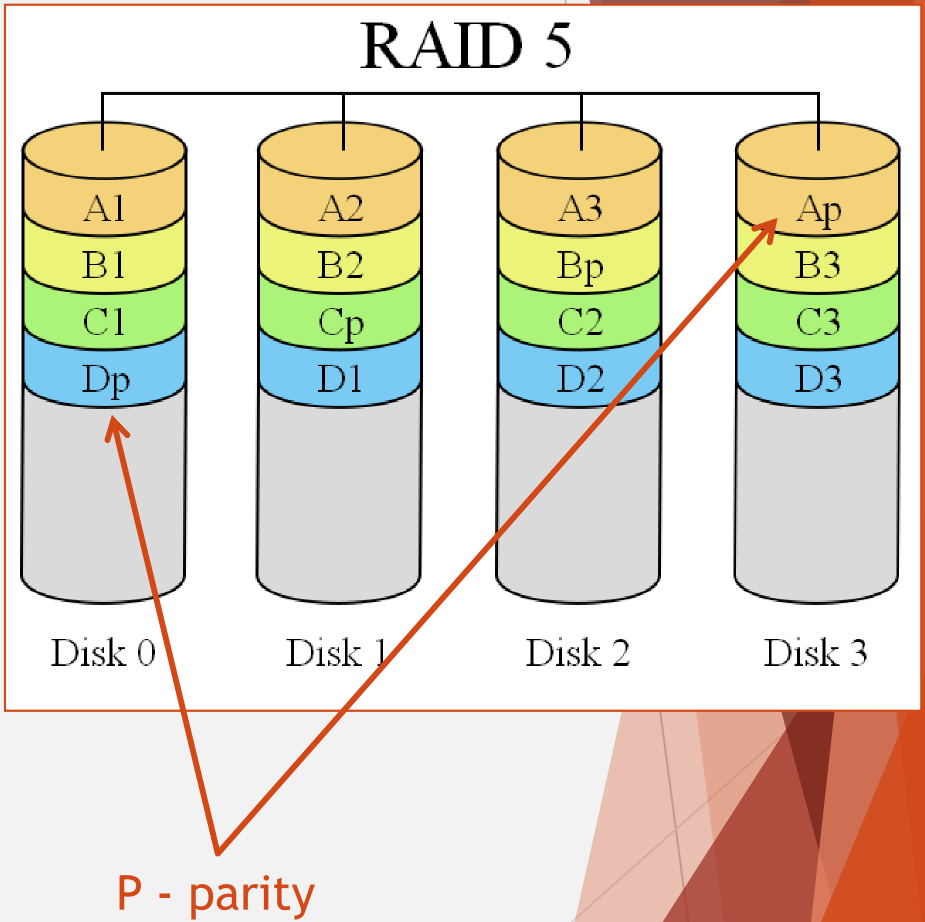
RAID 5

-**RAID 5** is an ideal combination of good performance, good fault tolerance and high capacity and storage efficiency.

- An arrangement of parity and CRC to help rebuilding drive data in case of disk failures.

- Uses a block of data for multiple disks, together with the distribution of the parity bits across the matrix structure.

-**“Distributed Parity”** is the key word here.

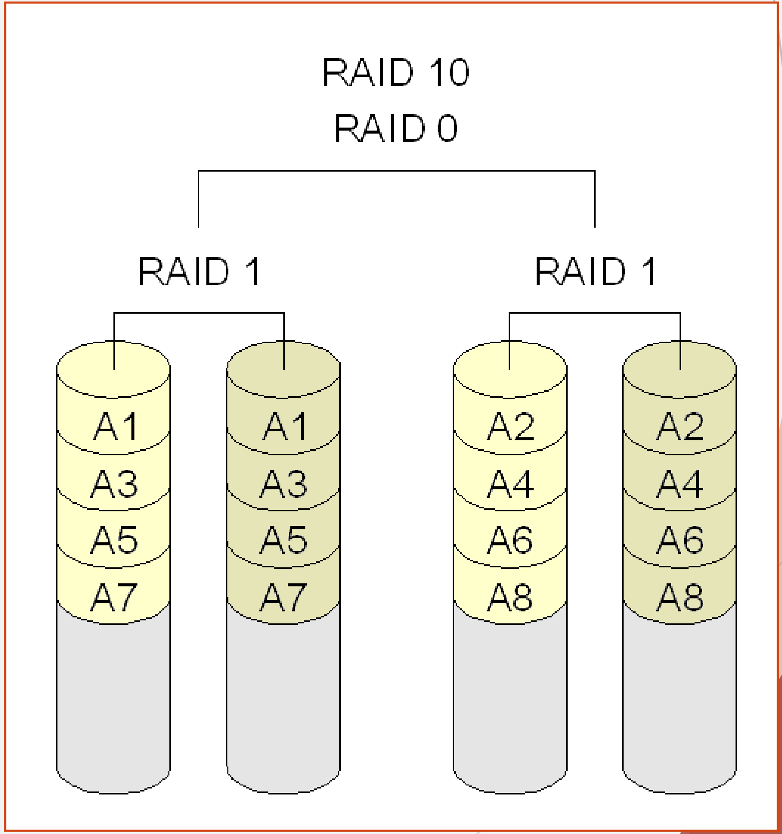


RAID 10

- Combines **RAID 1** and **RAID 0**.

- Which means having the pleasure of both -good performance and good failover handling.

- Also called ‘Nested RAID’.



Software Based RAID

- Software implementations are provided by many Operating Systems.

- A software layer sits above the disk device drivers and provides an abstraction layer between the logical drives(**RAID**s) and physical drives.

- Server's processor is used to run the **RAID** software.

- Used for simpler configurations like **RAID 0** and **RAID 1**.

References:

1. David A. Patterson, Garth Gibson and Randy H. Katz. A Case for Redundant Arrays of Inexpensive Disks (RAID) (URL: https://web.mit.edu/6.033/2015/wwwdocs/papers/Patterson88.pdf)

2. Attempting to restore this backup may encounter storage space problems & Exclusive access Error

(URL: https://dba.stackexchange.com/questions/154398/attempting-to-restore-this-backup-may-encounter-storage-space-problems-exclusive-access-Error)

3. Урбанович, П. П. Информационная безопасность и надежность систем : учебно-методическое пособие по одноименному курсу для студентов специальности 1-40 01 02-03 "Информационные системы и технологии" / П. П. Урбанович, Д. М. Романенко, Е. В. Романцевич. - Минск : БГТУ, 2007. - 87 с. (URL: http://elib.belstu.by/handle/123456789/2937)

4. Урбанович, П. П. Защита информации методами криптографии, стеганографии и обфускации : учеб.-метод. пособие для студ.. - Минск : БГТУ, 2016. - 220 с. (URL: http://elib.belstu.by/handle/123456789/23763)

5. Урбанович, П. П. Защита информации и надежность информационных систем : пос. для студ. вузов спец. 1-40 05 01-03 «Информационные системы и технологии (издательско-полиграфический комплекс)» / П. П. Урбанович, Д. В. Шиман.- Минск : БГТУ, 2014. - 91 с. (URL: https://elib.belstu.by/handle/123456789/23761)

6. Ochrona informacji w sieciach komputerowych / pod red. prof. P. Urbanowicza. - Lublin: Wydawnictwo KUL, 2004. - 150 p. (URL: https://elib.belstu.by/handle/123456789/27516)

7. Patsei, N. V. Information scrambler/descrambler based on combination of data compression and error-correction codes/N. V Patsei, P. P. Urbanovich// II Intern. Symposium Proceedings “New Electrical and Electronic technologies and their Industrial Implementation“ – Poland, Kazimierz Dolny, 14-17.02, 2001. – P.78-80. (URL: https://elib.belstu.by/handle/123456789/26755)

8. Urbanovich, P. P. Failure Consequences Imitation Modeling in the Equipment of Data Protection and Data Store Systems / P. P. Urbanovich, A. I. Halaburda, D. M. Romanenko, S. B. Makas // III Intern. Symposium NEET'2003. Zakopane, Poland, 13-16.05, 2003.- P.210-212. (URL: https://elib.belstu.by/handle/123456789/26742)

9. Использование многоуровневых турбо кодов в системах передачи двоичных данных / П. П. Урбанович [и др.] // Автоматический контроль и автоматизация производственных процессов : материалы Международной научно-технической конференции, 6-8 июня 2006 г. - Минск, 2006. - С. 206-208. (URL: https://elib.belstu.by/handle/123456789/25727)

10. Исследование основных характеристик многоуровневых турбо кодов / П.П. Урбанович [и др.] // Автоматический контроль и автоматизация производственных процессов: материалы Международной научно-технической конференции, 6-8 июня 2006 г., Минск. – Мн.: БГТУ, 2006. – С. 199-202. (URL: https://elib.belstu.by/handle/123456789/25728)

11. [Electronic Resource], http://searchstorage.techtarget.com/generic/0,295582,sid5_gci1196310,00.html

13. [Electronic Resource], http://en.wikipedia.org/wiki/RAID_0#RAID_0

14. [Electronic Resource], http://en.wikipedia.org/wiki/RAID

15. [Electronic Resource], http://www.lascon.co.uk/d008005.html

16. Урбанович, П. П. Избыточность в полупроводниковых интегральных микросхемах памяти / П. П. Урбанович, В. Ф. Алексеев, Е. А. Верниковский. - Минск : Навука і тэхніка, 1995. - 262 с. (URL: https://elib.belstu.by/handle/123456789/24777)