Estonian Information Technology College

Disk management

Operating systems I800

Edmund Laugasson
edmund.laugasson@itcollege.ee

There has been used materials from Margus Ernits, Katrin Loodus when creating current slides.

Operating system external memory

- hard disks
- disk arrays
- Interchangeable discs
  - CD, DVD, BD etc
  - etc
- *Flash* and other memory devices

Disk structure

- track
- cylinder
- sector
- head
- *CHS (cylinder, head, sector)*
- *LBA Logical Block Addressing*
  - every block has its own address
  - LBA address =c*H*S+h*S+(s-1)
    - *c – cylinder, H – heads per cylinder, S – sectors per track, s - sector*
- how rotating HDD works (videos)
  - https://en.wikipedia.org/wiki/File:Harddrive-engineerguy.ogv
  - https://en.wikipedia.org/wiki/File:HardDisk1.ogg

about SSD
- how it works
- video1
- video2

HDD (Hard Disk Drive) parameters

- Volume
  - 1 GB (GigaByte) (10^9) and 1 GiB (GibiByte) (2^30)
  - HDD manufacturers are using GB = 1000 MB
  - technically they are correct
- the quantity of tracks, cylinders, heads and sector size
- speed
  - rotating speed
    - *revolutions per minute* (rpm)
    - 5400, 7200, 10000, 15000

http://www.storagereview.com/ssd_vs_hdd

HDD parameters 2

- speed (continues)
  - read, write speed
    - random block reading (important seek time)
    - linear reading (important fast rotating speed)
  - *seek time* – time spent to find exact block
  - *access time* – time spent to access certain block
  - *latency* – „wasted time“
- Reliability
  - MTBF *mean time between failures* (in hours)
    average working time without failures
  - S.M.A.R.T. *Self-Monitoring, Analysis and Reporting Technology*
    https://help.ubuntu.com/community/Smartmontools
    https://en.wikipedia.org/wiki/Lm_sensors - *Linux-monitoring sensors*

Naming disks in system

- UNIX-like and Linux
  - SCSI, SATA, SAS /dev/sd*
    - first SCSI disk /dev/sda
    - second SCSI disk /dev/sdb
  - IDE /dev/hd*
    - first IDE master disk /dev/hda
    - second IDE slave disk /dev/hdb
    - third IDE master disk /dev/hdc
- MS Windows
  - A:, B:, C:, D: etc disks
- Mac
  - /dev/disk0s1

Naming disks in system 2

- to avoid confusion there are alternative naming options
  - /dev/sd*
  - UUID = 5f473d92-42fd-459f-8bbc-3e52cbf1d368
    - GRUB (*/boot/grub/grub.cfg*), */etc/fstab* are using (**NB! Ensure which file can be changed manually!**)
    - blkid (block devices ID's)
    - lsblk (list block devices)
  - give names for partitions like LABEL = backup

Naming disks in Linux

- ls -l /dev/sda
  - brw-rw---- 1 root disk 8, 0 2008-02-24 03:26 /dev/sda
  - access rights: owner (root), group (disk)
  - device class number (*major device no*) 8 (SCSI devices)
  - device number (*minor device no*) 0 (first device)
- /dev/sda first SCSI disk
- /dev/sdb second SCSI disk
- /dev/hda first channel *master* disk
- /dev/hdb first channel *slave* disk
- /dev/hdc second channel *master* disk
- http://tldp.org/HOWTO/Partition/devices.html

Naming disks in Linux 2

- device classes
  - important to read for *driver* developers (see also http://www.linux-drivers.org/, https://www.kernel.org/doc/)
  - *Linux Assigned Names and Numbers Authority* (LANANA), *Linux Device List*
  - https://en.wikipedia.org/wiki/Linux_Assigned_Names_and_Numbers_Authority
  - http://www.lanana.org/docs/device-list/ (search: *8 block*)
  - https://www.kernel.org/pub/linux/docs/lanana/device-list/
- devices – in UNIX-like (incl Linuxis) device is a file, https://en.wikipedia.org/wiki/Device_file
  - **c** *character device* unbuffered direct access to hardware (in Linux a virtual device generally)
  - **b** *block device* buffered access to hardware (in Linux a physical device generally)
  - http://stackoverflow.com/questions/1823743/knowing-a-device-special-file-major-and-minor-numbers-in-linux

Partitions

- disk can be splitted into partitions
- partition table goes into 0-sector, MBR (Master Boot Record)
  - up to 446 B boot loader (e.g. GNU GRUB)
  - 512 B together with partition table
  - cylinder 0, track 0, sector 0
  - EST https://wiki.itcollege.ee/index.php/MBR
  - ENG https://en.wikipedia.org/wiki/Master_boot_record
- Nowadays GPT (EFI), where partition table is directly written onto partition and not in separate, special place
  - *with Intel Itanium* processor machines
  - no quantity limit, size >2TB allowed
  - EST https://wiki.itcollege.ee/index.php/GPT
  - ENG https://en.wikipedia.org/wiki/GUID_Partition_Table

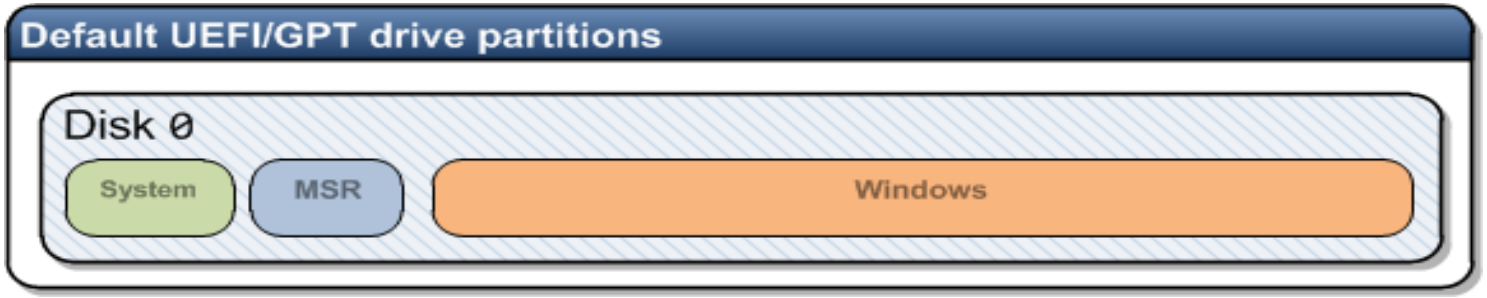
Partitions 2

- there was a belief that 4 partitions would be enough
  - *primary partition*, up to 4 (MBR, also DOS partition table)
    - disk names between 1-4 (/dev/sda1...sda4)
  - *extended partition*,
    one extended partition per disk is allowed
    - disk names between 1-3 (/dev/sda1...sda3)
    - extended partition with at least one logical volume
      - disk names between 5 - * (/dev/sda5...sda*)
    - *lsblk* (to view, *man lsblk*), http://linux.die.net/man/8/lsblk
    - *df -h* (free disks space, *man df*), https://linux.die.net/man/1/df

Partitions in MS Windows

- Understanding Disk Partitions
  https://technet.microsoft.com/en-us/library/dd799232(v=ws.10).aspx
- MS Windows and GPT FAQ
- https://msdn.microsoft.com/en-us/library/windows/hardware/dn640535(v=vs.85).aspx
- EST https://wiki.itcollege.ee/index.php/DISKPART
- ENG
  https://technet.microsoft.com/en-us/library/cc770877(v=ws.11).aspx

  by default based on MS Windows 7 and MS Windows Server 2008 R2 GPT disk partitioning table:
  MSR -Microsoft® Reserved Partition



creating partitions

- disk information
  - fdisk -l [disk] e.g.: **fdisk -l /dev/sda**
  - Disk /dev/sda: 80.0 GB, 80026361856 bytes
  - 255 heads, 63 sectors/track, 9729 cylinders
  - Units = silindrit of 16065 * 512 = 8225280 bytes
  - Disk identifier: 0xc51bc51b
  - Device Boot Start End Blocks Id System
  - /dev/sda1 * 1 5571 44749026 7 HPFS/NTFS
  - /dev/sda2 5572 9552 31977382+ 83 Linux
- tools to create partitions
  - fdisk
  - cfdisk
  - parted (suggested to use in case of GPT partition table)
  - https://en.wikipedia.org/wiki/List_of_disk_partitioning_software

using fdisk

- fdisk /dev/sd*
- **m** *manual*
- **p** displays partition table (*print*)
- **n** creates a new partition (*new*)
- **d** deletes a partition (*delete)*
- **q** will quit without saving partition table (*quit)*
- **w** writes partition table onto HDD (*write*)

creating a file system

- before you can use the partition, a file system must be created
  - ENG https://en.wikipedia.org/wiki/File_system
  - EST https://wiki.itcollege.ee/index.php/Failisüsteem
- ***NB!*** Please pay attention to symbols how to mark disks compared with fdisk
- mkfs -t type partition
  - mkfs -t ext4 /dev/sdb*
  - mkfs.ext4 /dev/sdb*

mounting partitions

- the ***mount*** command is used
- mount -t <type> <parameters> <device> <full directory path as a mount point>
  - **mount** -t ext4 -o rw /dev/sda2 /home
    - will mount first SCSI device under the directory /home
  - **mount** -t ntfs /dev/sda1 /mnt/ntfs
    - will mount in read-only mode NTFS partition sda1 under the directory /mnt/ntfs
  - **ntfs-3g** – 3rd generation NTFS driver for reading and writing
    - mount -t ntfs-3g /dev/sda1 /mnt/ntfs/
    - ntfs-3g /dev/sda1 /mnt/ntfs
- unmount
  - **umount** /home
- ENG
  - man mount, man ntfs-3g, etc
- EST
  - https://wiki.itcollege.ee/index.php/Mount
  - https://wiki.itcollege.ee/index.php/NTFS_vs_Ext4
  - https://wiki.itcollege.ee/index.php/NTFS_failis%C3%BCsteemi_kasutamine_linuxis_v%C3%B5i_macis

mount partitions at boot: */etc/fstab*

- man fstab, EST https://wiki.itcollege.ee/index.php/Fstab
- https://guidgenerator.com/ - UUID (GUID)

```
$ cat /etc/fstab
# /etc/fstab: static file system information.
#
# Use 'blkid' to print the universally unique identifier for a
# device; this may be used with UUID= as a more robust way to name devices
# that works even if disks are added and removed. See fstab(5).
#
# <file system> <mount point>   <type>  <options>       <dump>  <pass>
# / was on /dev/sda5 during installation
UUID=da17263a-2afb-42c9-86fb-abe4f55c12f4 /    ext4    errors=remount-ro,noatime,commit=600 0
1
# /home was on /dev/sdb1 during installation
UUID=ba77de4a-0a35-4968-bf97-d93840165904 /home   ext4    defaults,noatime,commit=600        0
2
# swap was on /dev/sda1 during installation
#UUID=68682f78-7849-46c5-a2fa-125f3ee72e99 none     swap    sw              0       0
/dev/mapper/cryptswap1 none swap sw 0 0
#/dev/disk/by-uuid/7fa84b49-ae7c-4a64-be4e-04c22a32ade0 /mnt/7fa84b49-ae7c-4a64-be4e-04c22a32ade0
auto nosuid,nodev,nofail,x-gvfs-show,errors=remount-ro,noatime,commit=600 0 0
/dev/disk/by-uuid/6cfbe92f-78f7-4eb9-ab17-fe09267a1bb5 /mnt/data2 auto
nosuid,nodev,nofail,noauto,errors=remount-ro,noatime,commit=600 0 0
```

MS Windows'is UUID:
https://msdn.microsoft.com/en-us/library/windows/desktop/aa373928(v=vs.85).aspx

/etc/fstab

- UUID=6...703e / ext4 defaults,acl,errors=remount-ro 0 1
- device
  - can be as /dev/sda1
  - can be as UUID=
  - can be as LABEL=
- mount point
  - where the device will be mounted
    - e.g. / root directory
    - e.g. */home* for user home directories (useful as there is no need to backup user data when new system installation needed)

/etc/fstab (2)

- UUID=6...703e / ext4 defaults,acl,errors=remount-ro 0 1
- file system type
  - e.g. ext3, ext4, reiserfs, btrfs, ntfs, cifs etc
- Options
  - *sync/async* – synchronous or asynchronous I/O
  - *noauto* – file system will be not mounted during boot

/etc/fstab (3)

- UUID=6...703e / ext4 defaults,acl,errors=remount-ro 0 1
- *noexec* – running binary executables prohibited
- *suid* – *suid* bit allowed
- *ro* – will be mounted as read-only
- *user* – user can mount the file system
- *defaults* – default settings will be used (differs by file system)
  - *rw, suid, dev, exec, auto, nouser, async*
- *acl* – the Access Control List will be allowed

File system check

- File systems may be broken sometimes
  - e.g. pull out a memory stick without syncing in first place
    - use sync command in CLI
  - current passes away
- **fsck [parameters] -t <type> <device>**
- also fsck.* commands can be used
  - fsck.ext4 -p /dev/sda1

Restore partitions

- gpart /dev/sdX
  - takes time but guesses partitions and displays the table
  - not installed by default
- gpart -W /dev/sdX /dev/sdX
  - when table seems to be logical then it will be written onto disk
- it would be a good idea to make a backup from your partition table and keep it in safe and secure place
  - dd if=/dev/sda of=sda.mbr count=1 bs=512
  - sfdisk -d /dev/sda > sda.tabel
- to restore:
  - dd if=sda.mbr of=/dev/sda
  - sfdisk /dev/sda < sda.tabel

Swapping

- Swap area is used to extend RAM on HDD
- makes RAM usage more flexible
- virtual address space
  - physical aadress
  - MMU memory management unit - will translate virtual memory address into physical
  - virtual address
  - memory page
- swapping
  - swap in/swap out

SWAP (2)

- when any of memory part is not used then corresponding memory blocks are written to swap area
- swap can be separate partition or file on the partition
  - MS Windows:
    would be good idea to put onto separate partition
- create: *mkswap /dev/sdb1*
- activate: *swapon /dev/sdb1*
- allow on boot in file */etc/fstab*
  - `/dev/sdb1 none swap sw 0 0`
  - `UUID=<hash> none swap sw 0 0`

SWAP (3)

- creating a swap file
  - *dd if=/dev/zero of=/swapfile bs=1024 count=1048576*
  - will create a 1GiB file with zeros inside
  - create: *mkswap /swapfile*
  - activate: *swapon /swapfile*

SWAP (4)

- swap size
  - in old times there was a thumb rule 1-2 times of RAM size
  - nowadays can be less
  - on wearable computers it would be useful to create ⩾RAM size as *hibernate* will use swap area (in case of Linux)
  - in case of SSD disk the swap area:
    - ensure that TRIM works
      - https://askubuntu.com/questions/464306/a-command-which-checks-that-trim-is-working
    - there is suggested to avoid *hibernate*
    - using *RAM-drive* and avoiding swap area, -file
    - SSD
      - ENG https://en.wikipedia.org/wiki/Solid-state_drive
      - EST https://wiki.itcollege.ee/index.php/SSD_kettad
    - *https://en.wikipedia.org/wiki/RAM_drive*
    - *https://en.wikipedia.org/wiki/List_of_RAM_drive_software*

Problem

- using disks, several issues will rise
  - perfomance – increasing disk speed and size is not unlimited. Fast disks are expensive
  - Latency – disks with small latency are expensive
  - Reliability – hard disks have MTBF (mean time between failures). Quite often the disk will failure in first place.
  - ***Planning disk space is quite challenging*** – sometimes is needed more, sometimes less

Problem solving

- Pefomance
  - write and read in parallel from multiple disks
- Latency
  - read and write in parallel from multiple disks
- Reliability
  - create redundancy and duplicate data on multiple disks
- Planning and partitioning
  - use the system that allows change partition size

RAID

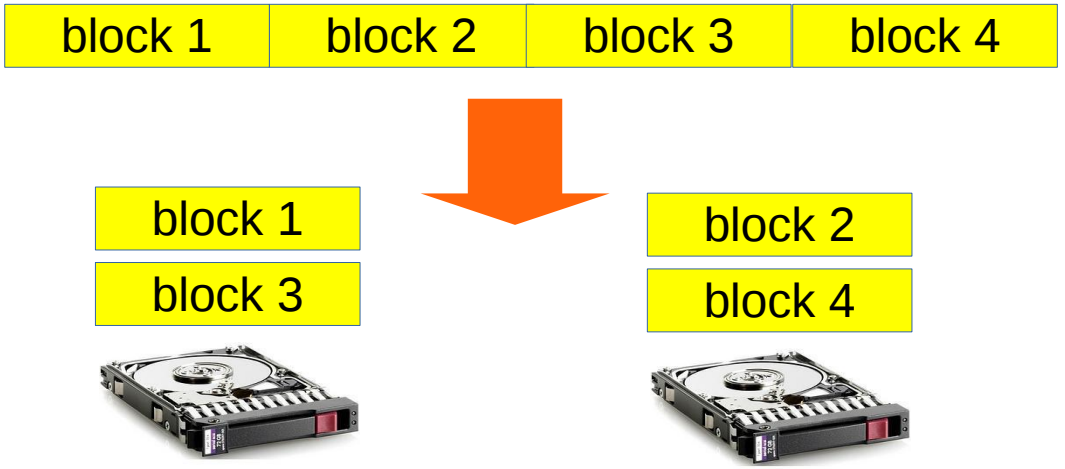
- RAID *Redundant Arrays of Independent Disks*
- *Redundancy* is used to increase reliability and perfomance
- *Mean Time Between Failures* (MTBF)
  - measured in hours
- ENG https://en.wikipedia.org/wiki/RAID , https://en.wikipedia.org/wiki/Category:RAID
- EST https://wiki.itcollege.ee/index.php/RAID
- EST https://wiki.itcollege.ee/index.php/Raid_kettas%C3%BCsteemid
- EST https://et.wikipedia.org/wiki/S%C3%B5ltumatute_ketaste_liiasmassiiv
- ENG https://en.wikipedia.org/wiki/Mean_time_between_failures

RAID

- Stripe data storage – data blocks are stored across over multiple disks
- Redundancy – data is duplicated on multiple disks
  - disk *mirroring*
- Error Correction Code – there will be checksums calculated that can be used to discover errors and in some cases also fix
- Rebuild – restore disk based on information at other disks
- Spare disk – a backup disk connected to the system

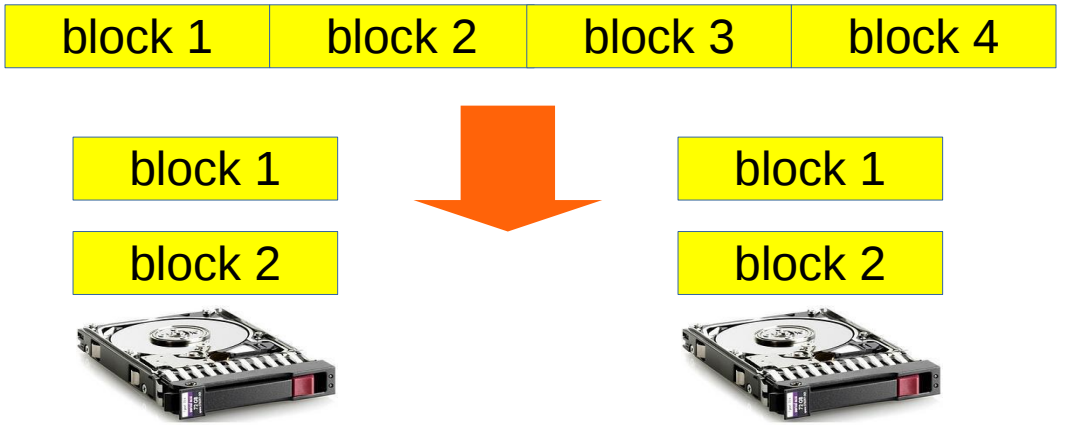
RAID 0

- RAID 0
  - data stored as *striped* on multiple disks
  - increases read - and write speed
  - reliability does not increase but reduce
  - Min disks: 2. Data storage volume is a sum of disks



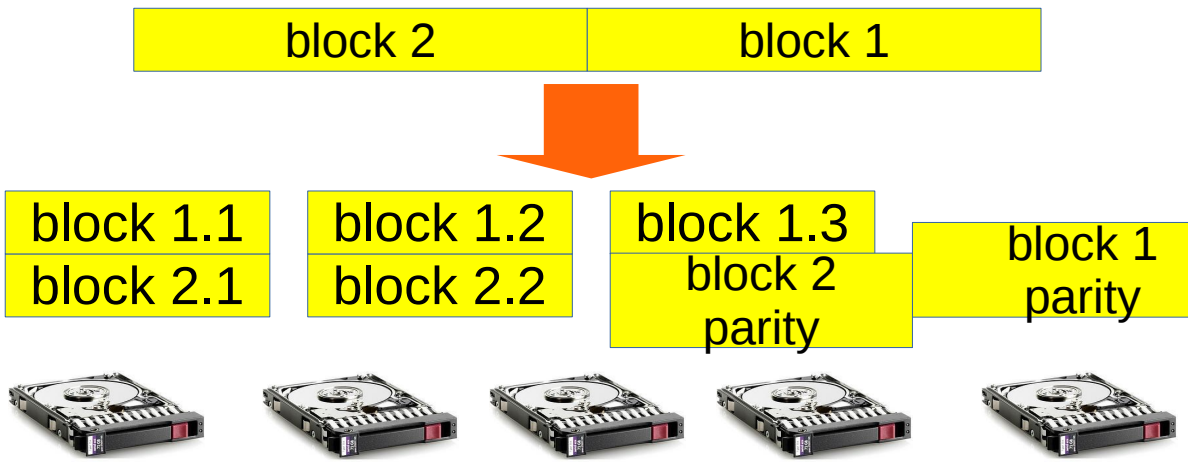
RAID 1

- RAID 1
  - disks mirroring or creating a redundancy
  - reliability will be increased because data is duplicated
  - reading speed will be increased as data can be read from multiple disks at the same time
  - writing speed is the same as data will be written onto multiple disks
  - Min 2 disks. Volume = disks volume/2



RAID 5

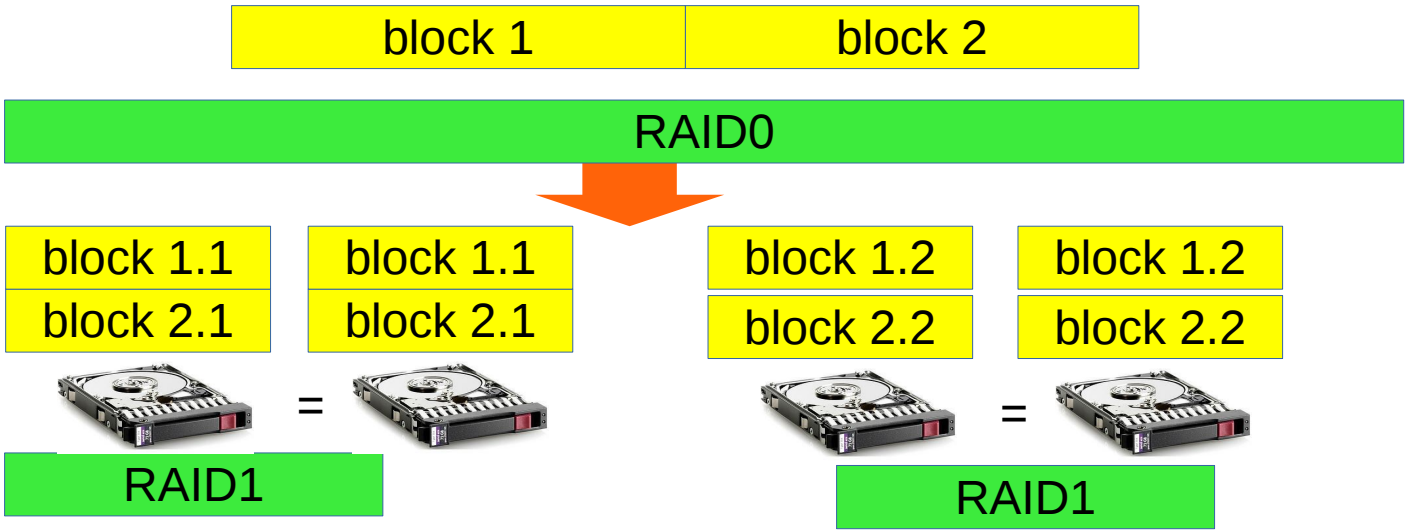
- RAID 5
  - stripe storage disks
  - shared parity check (parity information is shared between disks)
  - damaged disk data will be restored due to parity information on other disks
  - creating parity information takes a time



large reading speed
average writing speed
slow array restore
file -, mail -, media server
most widely used

RAID 1+0

- RAID 1+0
  - multiple mirrors in stripe storage
  - combines RAID0 and RAID1 good characteristics (redundancy and speed)
  - used in case of database servers and other high perfomance apps
  - high price



rich world of RAID

- Nested RAID:
  - RAID0+1
  - RAID5+0
  - RAID6+0+0 jne
- intermediate levels
  - RAID2, stripe storage on bit level
  - RAID3, stripe storage on byte level
  - RAID4, stripe storage on block level (parity information on separate disk)
  - http://www.acnc.com/raid

RAID controller

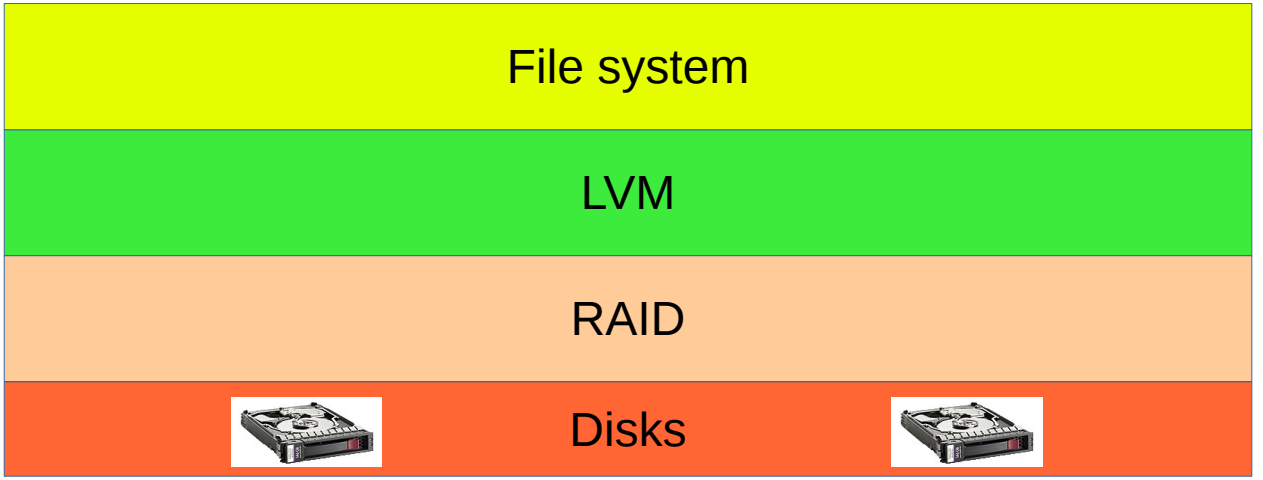
- manages I/O queries
- restores data from spare drives when needed
- monitor errors
- hardware and software RAID-controller
- ENG https://en.wikipedia.org/wiki/Disk_array_controller
- EST https://wiki.itcollege.ee/index.php/RAID_controller
- ENG http://www.tldp.org/HOWTO/Software-RAID-HOWTO.html
- EST https://wiki.itcollege.ee/index.php/Softi_RAID_Ubuntu_baasil.
- ENG https://help.ubuntu.com/community/Installation/SoftwareRAID
- EST https://wiki.itcollege.ee/index.php/Tarkvaraline_RAID_Ubuntus

RAID summary

- increases reliability and/or speed
- is more expensive than regular storage
- will increase complexity that may cause problems
- Will not replace the need to do backup
- allows renew system to breake mirroring so the one disk will keep original system
  - when there is a success of renewing the system then new system will be mirrored to another disk after rebuilding RAID
  - when renewing system caused problems then data will be mirrored back from old drive

LVM

- Logical Volume Manager
- created as logical layer between disks and sile system
  - flexible way to configure physical disks
  - widely used in large amount of disks deployments

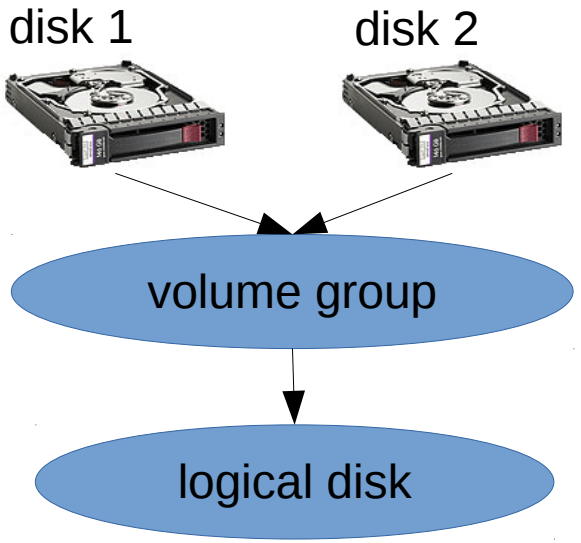


LVM

- allows
  - add new disks to the system by enlarging existing file system (e.g. adding additional storage space for home folders)
  - remove old disks from the system
    - data can be easily copied to new disks and unmount old ones
  - *Snapshot copy*
  - can be combined with RAID usage

LVM

- block device
  - hard disks or partitions
- disk groups or *volume groups*
- logical disks



LVM usage steps

- pv* - physical volume commands
- vg* - volume group commands
- lv* - logical volume commands
- initialize disks
- create a volume group and bind disks with that
- share logical disks (partitions)
- format logical disk with file system
- mount formatted logical disk

LVM

- to use LVM there should be disk or partition prepared as a physical volume (all commands as root user)
  - **pvcreate <disk1> <disk2> ...**
  - e.g. **pvcreate /dev/sdb1**
- created physical volumes can be viewed:
  - **pvdisplay**

LVM

- after preparing (marking) disks there should be created a volume group
  - **vgcrearte <group> <disk1> <disk2> ...**
  - e.g.: **vgcreate mysql /dev/sdb1**
  - here has been created a volume group named mysql by including device /dev/sdb1 (in group there can be 1....n partitions)
- created volume groups can be displayed:
  - **vgdisplay**

LVM

- **vgrename <oldgroup> <newgroup>**
  - e.g.: **vgrename mysql db**
- creating new logical volume using lvcreate
- **lvcreate -L 100G -n lv01 mysql**
  - Logical volume "lv01" created
- after creating logical volume there can be created a file system on that and mounted under certain directory and start using it
- mkreiserfs /dev/vg1/lv01
- will create a reiserfs file system on logical volume lv01

LVM

- extending: lvextend
  - **lvextend -L150G /dev/mysql/lv01**
- reducing lvreduce
  - **lvreduce -L1G /dev/mysql/lv01**
- removing physical disk from LVM array
  - **pvremove <disk1> <disk2> ...**
  - e.g.: pvremove /dev/sdb1 /dev/sdc1 /dev/sdd1 /dev/sde1
- there can be removed 1...n partitions

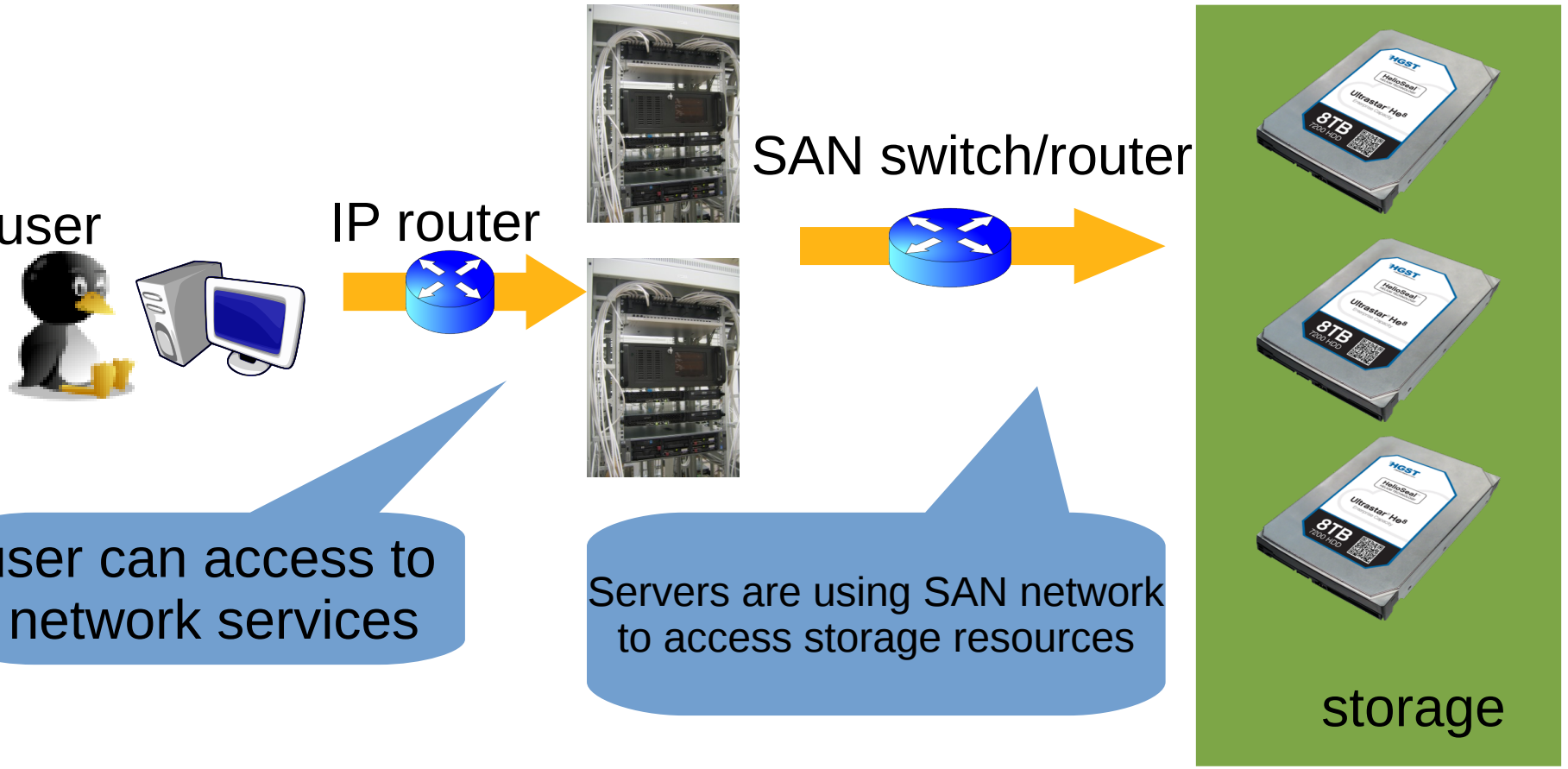
Disk quota

- sometimes you need to limit storage space given for users
- there will be used a disk quota for that
- **sudo apt-get install quota**
- when mounting, please use appropriate quota parameter
  - **mount -o quota /dev/sdb2 /home**
- for automatic mounting through /etc/fstab please use usrquota and/or *grpquota* parameter
  - **/dev/sda2 /home ext4 defaults,usrquota 1 2**

Disk quota

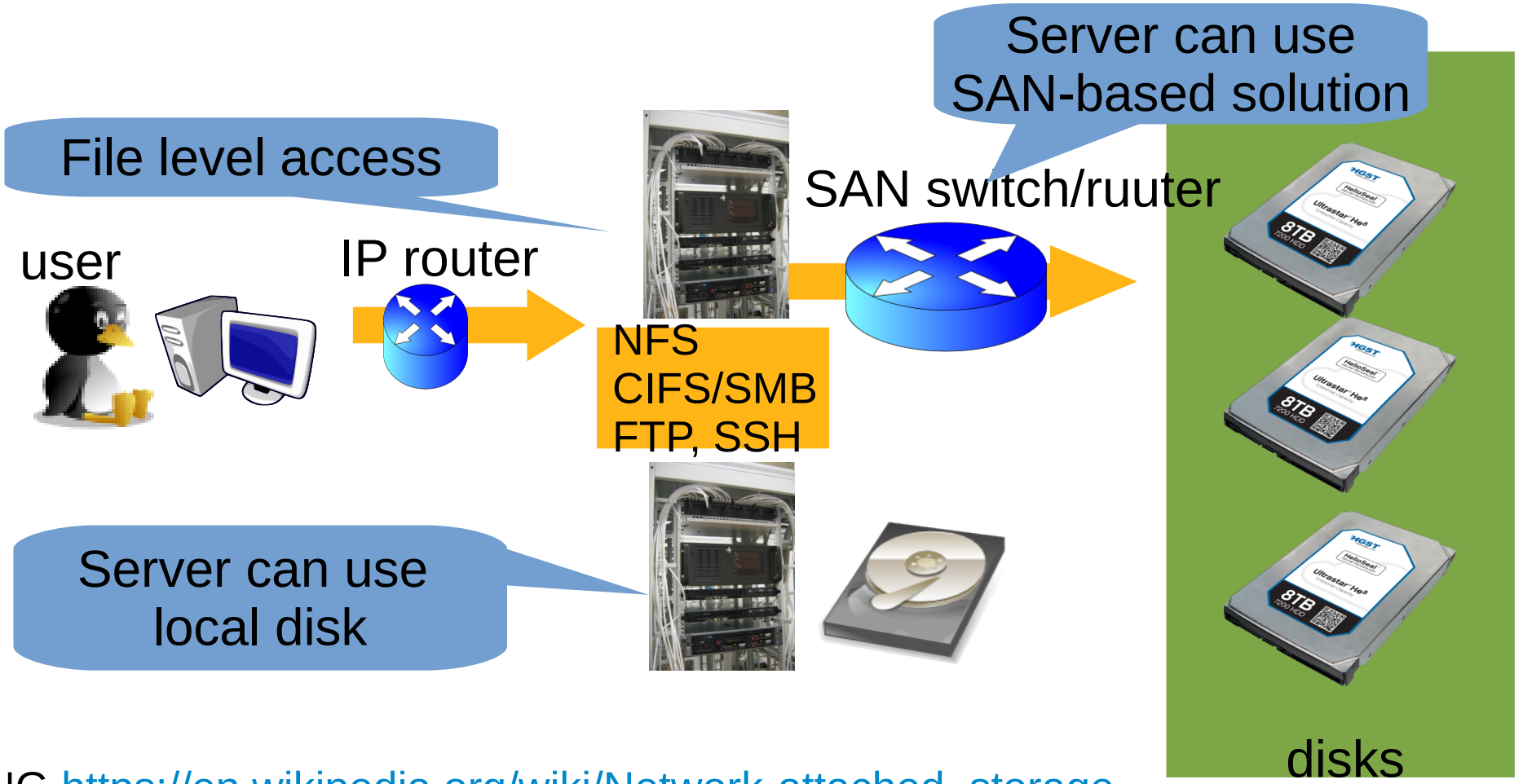
- to monitor disk quota limits
  - **quotacheck -avug**
- this would be useful to watch periodically after certain amount of time
- setting up quota for user and group
  - **edquota -u student**
  - **edquota -g students**
- soft limit – can be exceed temporarily
- hard limit – cannot be exceeded (data writing denied when reaching the limit)

Storage Area Network (SAN)



ENG https://en.wikipedia.org/wiki/Storage_area_network
EST https://wiki.itcollege.ee/index.php/SAN

Network Attached Storage (NAS)



ENG https://en.wikipedia.org/wiki/Network-attached_storage
EST https://wiki.itcollege.ee/index.php/NAS

Links

- https://help.ubuntu.com/community/PartitioningSchemes
- EST https://wiki.itcollege.ee/index.php/Partitsioonid
- EST https://wiki.itcollege.ee/index.php/Kettajaod
- http://manpages.ubuntu.com/manpages/xenial/man8/gdisk.8.html
- EST https://wiki.itcollege.ee/index.php/Gdisk
- https://en.wikipedia.org/wiki/Disk_partitioning
- https://en.wikipedia.org/wiki/Partition_type
- https://support.microsoft.com/en-us/kb/302873 - GUID FAQ, MS Windows

Links

- RAID WikiPedia http://en.wikipedia.org/wiki/RAID
- RAID flash demo
  http://www.acnc.com/raid.html
- SAN in Linux
  http://www.linuxjournal.com/magazine/use-linux-san-provider
  http://iscsitarget.sourceforge.net/
  https://wiki.itcollege.ee/index.php/ISCSI_target_ja_initiator
  http://www.openfiler.com/, vt ka alternatiive
- NAS solutions:
  - OpenMediaVault
    https://www.howtoforge.com/tutorial/install-open-media-vault-nas/
  - FreeNAS
    https://www.howtoforge.com/network_attached_storage_with_freenas

Links (LVM)

- LVM for beginners http://www.howtoforge.com/linux_lvm
- http://www.tldp.org/HOWTO/LVM-HOWTO/
- http://www.debian-administration.org/articles/410
- http://linuxdevcenter.com/pub/a/linux/2006/04/27/managing-disk-space-with-lvm.html?page=1
- http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/4/html/Cluster_Logical_Volume_Manager/

Links (RAID)

- Software RAID in Ubuntu
  https://help.ubuntu.com/community/Installation/SoftwareRAID
  https://wiki.ubuntu.com/Raid
- FakeRAID - https://help.ubuntu.com/community/FakeRaidHowto
- RAID in server https://help.ubuntu.com/lts/serverguide/advanced-installation.html
- RAID in Linux https://raid.wiki.kernel.org/index.php/RAID_setup
- RAID intro http://www.linux-mag.com/id/7924/
- Gnome Disks
  https://en.wikipedia.org/wiki/GNOME_Disks
  http://askubuntu.com/questions/500549/how-to-run-gnome-disk-utility
- http://www.tldp.org/HOWTO/Software-RAID-0.4x-HOWTO-2.html
- http://www.raid-calculator.com

Questions?

Thank you for your attention!