File System & Swap Area

2019 Winter Wheel Seminar
tink@

File System

Files

- User's view: Named sequence of bytes
- File system's view: Collection of disk blocks

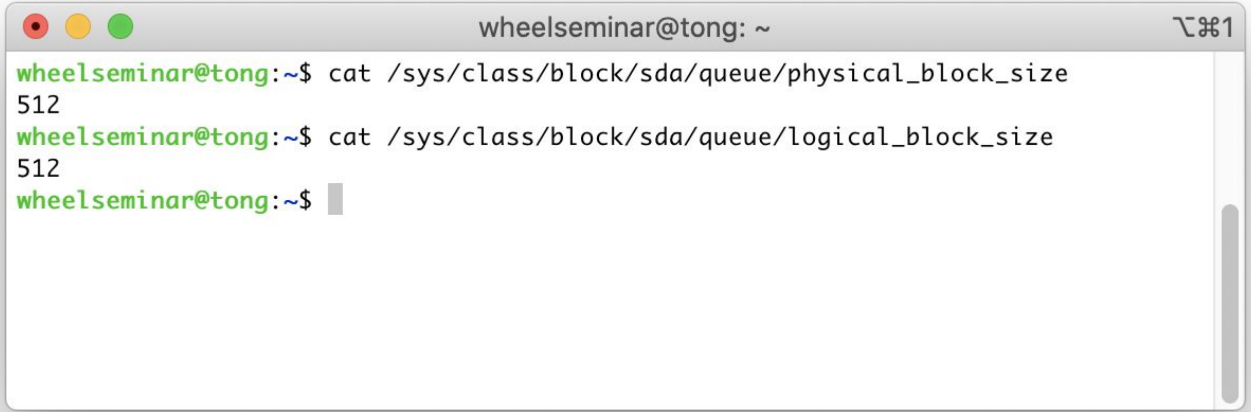
File System

- User's view: Manages files and data stored in files
- File system's view: Map name & offset to disk blocks

- Differs by OS

Block & Sector

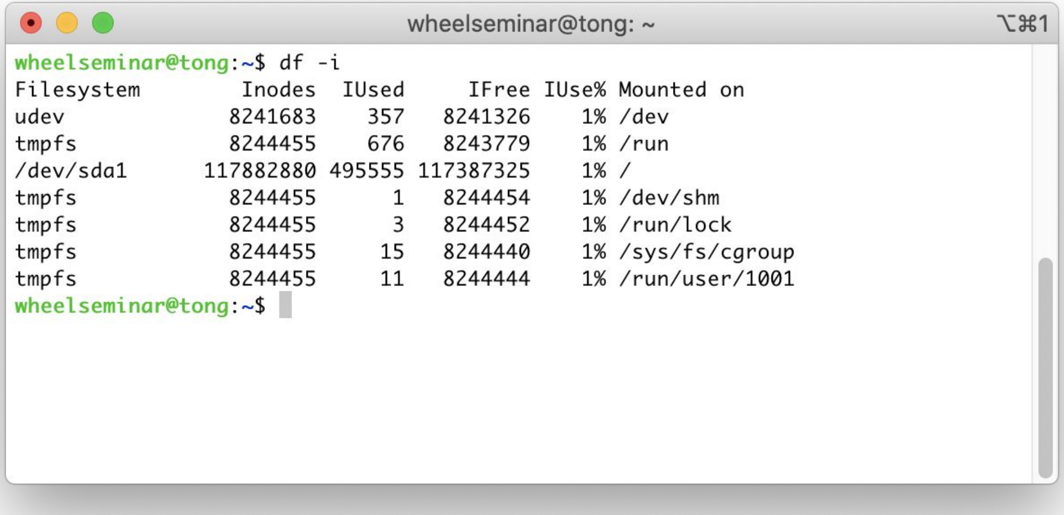
- Sector: Minimum storage unit of Hard Drive
- Block: Minimum unit for file system
  - multiple of sector size
  - Configurable



```
wheelseminar@tong:~$ cat /sys/class/block/sda/queue/physical_block_size
512
wheelseminar@tong:~$ cat /sys/class/block/sda/queue/logical_block_size
512
wheelseminar@tong:~$ 
```

Inode block & Data block

- Inode block
  - File Metadata (Data of data)
  - Position of data block
- Data block
  - Real file data



```
wheelseminar@tong:~$ df -i
Filesystem          Inodes  IUsed     IFree IUse% Mounted on
udev               8241683    357   8241326    1% /dev
tmpfs              8244455    676   8243779    1% /run
/dev/sda1        117882880 495555 117387325    1% /
tmpfs              8244455      1   8244454    1% /dev/shm
tmpfs              8244455      3   8244452    1% /run/lock
tmpfs              8244455     15   8244440    1% /sys/fs/cgroup
tmpfs              8244455     11   8244444    1% /run/user/1001
wheelseminar@tong:~$
```

Types of file systems

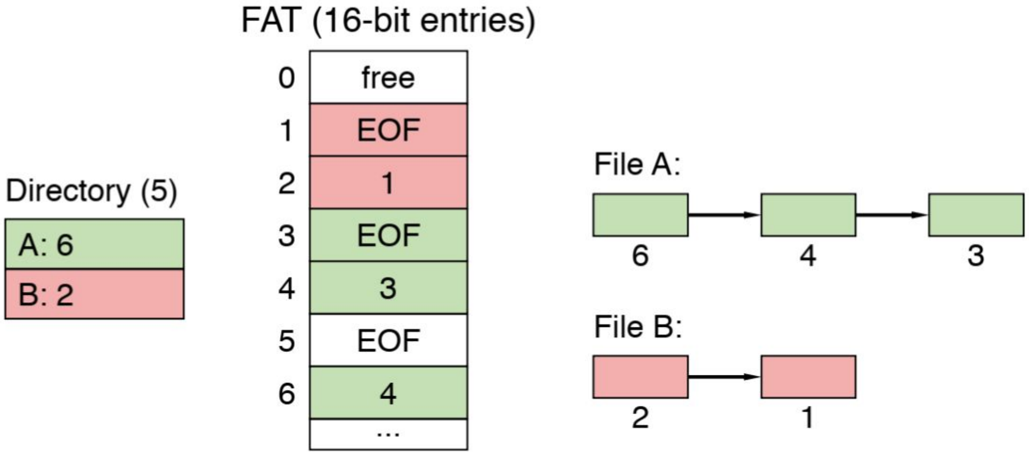
- Window: FAT16, FAT32, NTFS
- Linux: Btrfs, EXT2, EXT3, EXT4, ReiserFS, XFS
- MacOS: HFS+

Indexing Structure

- Linked files
- Indexed files

Linked files

- Linked list index structure
- File metadata (Inode) points file's first block
- File table: Linear map of all blocks on disk, each file is a linked list of blocks
- Example - Microsoft FAT (Linked files with cached pointers



FAT (16-bit entries)

0	free
1	EOF
2	1
3	EOF
4	3
5	EOF
6	4
	...

Image from: Youngjin Kwon CS330

Indexed files

- Each file metadata has an array holding all of its block pointers
- Random access is fast
- Max file size fixed by array's size => How to deal with this?

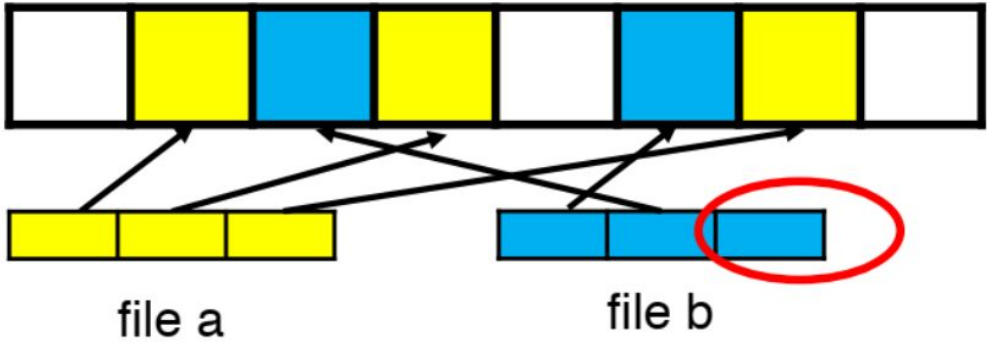


Image from: Youngjin Kwon CS330

Indexed files

- Max file size fixed by array's size => How to deal with this?

  => Multi-level indexed files

- Example: Berkely UNIX FFS

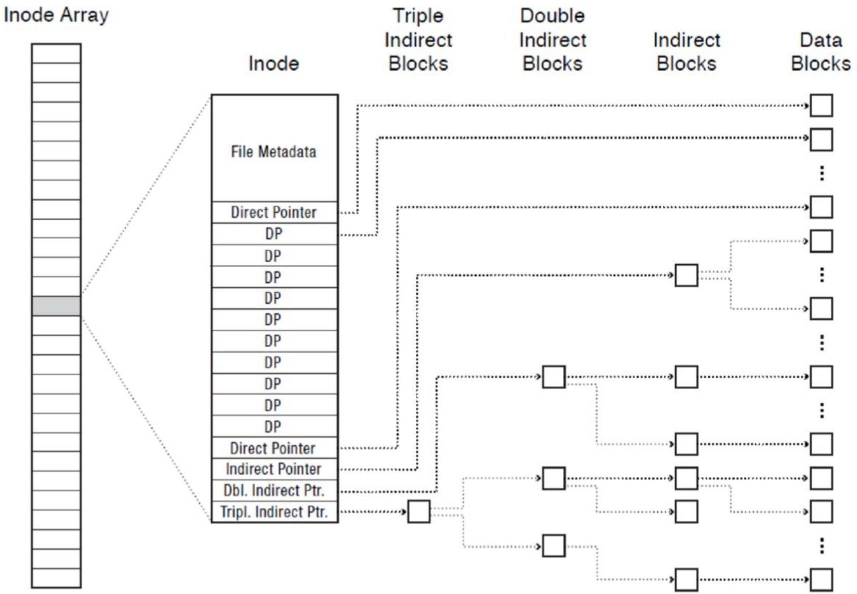


Image from: Youngjin Kwon CS330

Linux file systems

- EXT2
- EXT3 - Journaling
- EXT4 - Extents
- XFS - Journaling, Large on memory cache (Good performance)

Journaling

- Keeps track of changes not yet committed to the file system's main part by recording the intentions of such changes in a data structure
- Modes of Journaling (Has trade-offs)
  - Journal
  - Ordered
  - Write-back
- mount option "data=[mode]"

Extents

- Use contiguous area of storage reserved for a file
- Can store each range compactly as two numbers, instead of canonically storing every block number in the range
- Less file fragmentation

Managing file system in Linux

- Managing partition: fdisk
- Managing file system: mkfs
- Mount / Unmount device: mount/umount
- Check and restore file system: fsck

Partition

- Slice hard disk to one or more regions
- Each partition can be managed separately
  - Stable at problematic situations
  - Can format separately
- Types of partitions
  - Primary partition : Real disk partition
  - Extended / Logical partition : 1 disk - 1 extended partition - many logical partitions

Types of partition table layouts

- MBR
  - Max number of primary partitions is 4
  - Max size of partition is 2TB
- GPT
  - All partitions are primary partition

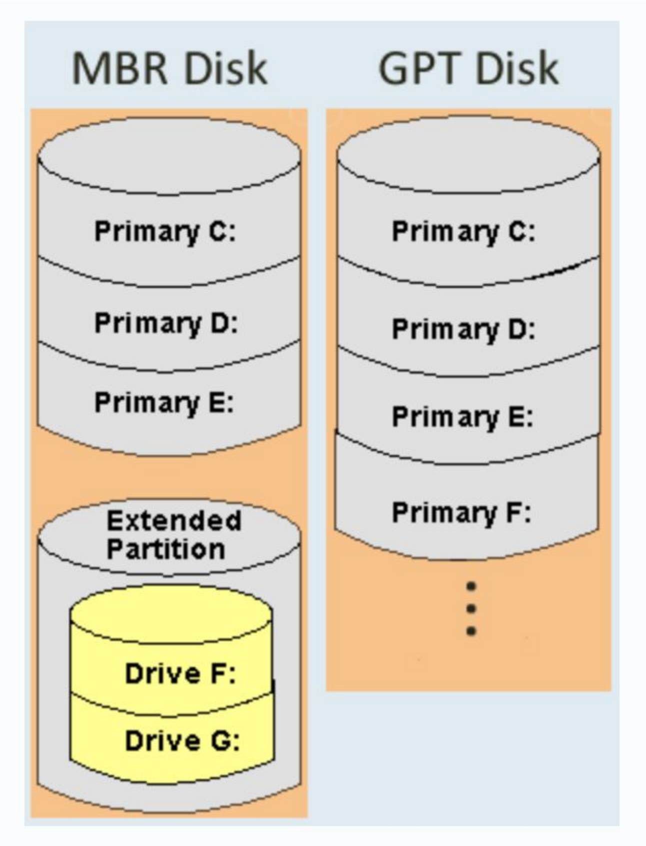
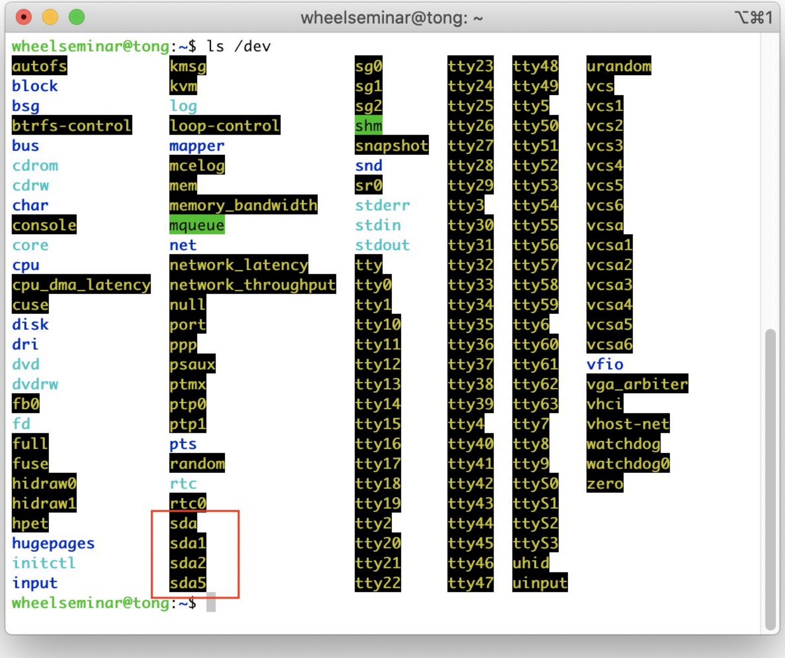


Image from: https://www.hdd-tool.com/nl/disk-converter/convert-primary-partition-to-logical.html

Managing partition in Linux

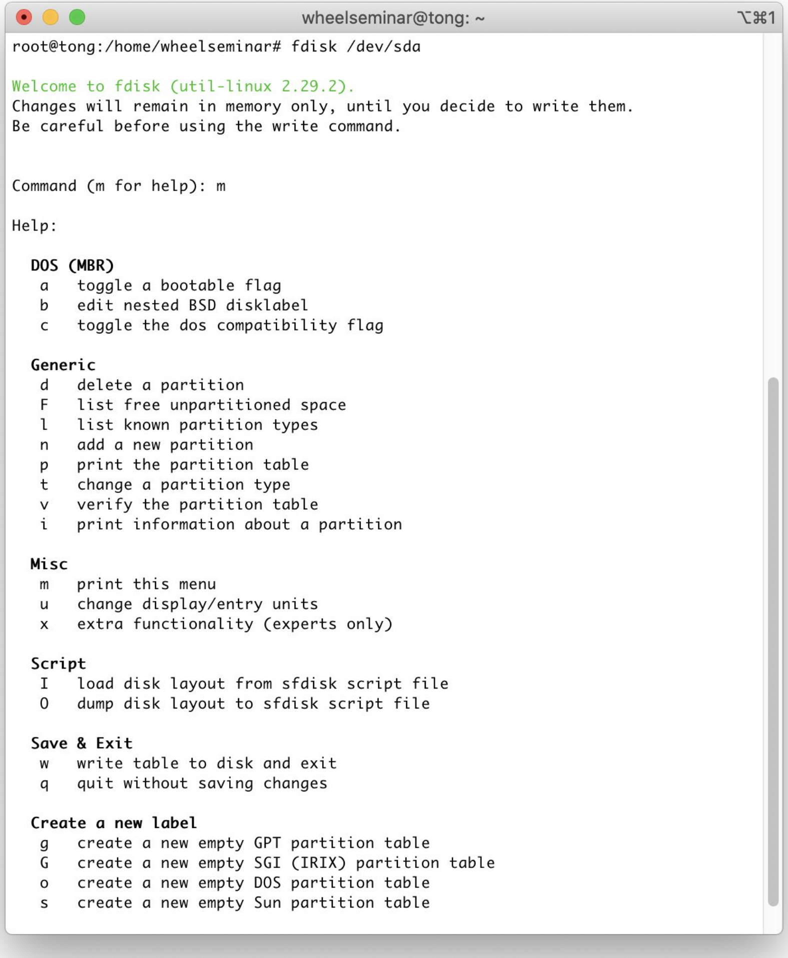
- fdisk, parted
- Linux system device files: /dev
  - IDE type hard disks: /dev/hda, /dev/hdb, /dev/hdc
  - Sata, schi type hard disks: /dev/sda, /dev/sdb, /dev/sdc



```
wheelseminar@tong:~$ ls /dev
autofs           kmsg                sg0       tty23  tty48   urandom
block            kvm                 sg1       tty24  tty49   vcs
bsg              log                 sg2       tty25  tty5    vcs1
btrfs-control    loop-control        shm       tty26  tty50   vcs2
bus              mapper              snapshot  tty27  tty51   vcs3
cdrom            mcelog              snd       tty28  tty52   vcs4
cdrw             mem                 sr0       tty29  tty53   vcs5
char             memory_bandwidth    stderr    tty3   tty54   vcs6
console          mqueue              stdin     tty30  tty55   vcsa
core             net                 stdout    tty31  tty56   vcsa1
cpu              network_latency     tty       tty32  tty57   vcsa2
cpu_dma_latency  network_throughput  tty0      tty33  tty58   vcsa3
cuse             null                tty1      tty34  tty59   vcsa4
disk             port                tty10     tty35  tty6    vcsa5
dri              ppp                 tty11     tty36  tty60   vcsa6
dvd              psaux               tty12     tty37  tty61   vfio
dvdrw            ptmx                tty13     tty38  tty62   vga_arbiter
fb0              ptp0                tty14     tty39  tty63   vhci
fd               ptp1                tty15     tty4   tty7    vhost-net
full             pts                 tty16     tty40  tty8    watchdog
fuse             random              tty17     tty41  tty9    watchdog0
hidraw0          rtc                 tty18     tty42  ttyS0   zero
hidraw1          rtc0                tty19     tty43  ttyS1
hpet             sda                 tty2      tty44  ttyS2
hugepages        sda1                tty20     tty45  ttyS3
initctl          sda2                tty21     tty46  uhid
input            sda5                tty22     tty47  uinput
wheelseminar@tong:~$ 
```

Managing partition in Linux: fdisk

- fdisk [disk device]
- Interactive
- No saving to disk before typing "w"



```
root@tong:/home/wheelseminar# fdisk /dev/sda

Welcome to fdisk (util-linux 2.29.2).
Changes will remain in memory only, until you decide to write them.
Be careful before using the write command.


Command (m for help): m

Help:

  DOS (MBR)
   a   toggle a bootable flag
   b   edit nested BSD disklabel
   c   toggle the dos compatibility flag

  Generic
   d   delete a partition
   F   list free unpartitioned space
   l   list known partition types
   n   add a new partition
   p   print the partition table
   t   change a partition type
   v   verify the partition table
   i   print information about a partition

  Misc
   m   print this menu
   u   change display/entry units
   x   extra functionality (experts only)

  Script
   I   load disk layout from sfdisk script file
   O   dump disk layout to sfdisk script file

  Save & Exit
   w   write table to disk and exit
   q   quit without saving changes

  Create a new label
   g   create a new empty GPT partition table
   G   create a new empty SGI (IRIX) partition table
   o   create a new empty DOS partition table
   s   create a new empty Sun partition table
```

Managing file system in Linux: mkfs

- Each file system has own mkfs command
  - example: mkfs.ext3
- mkfs is a front-end for this command
- mkfs [-c] [-t file_system_type] <partition_device>
  - -c option: bad sector test
  - -t default is ext2

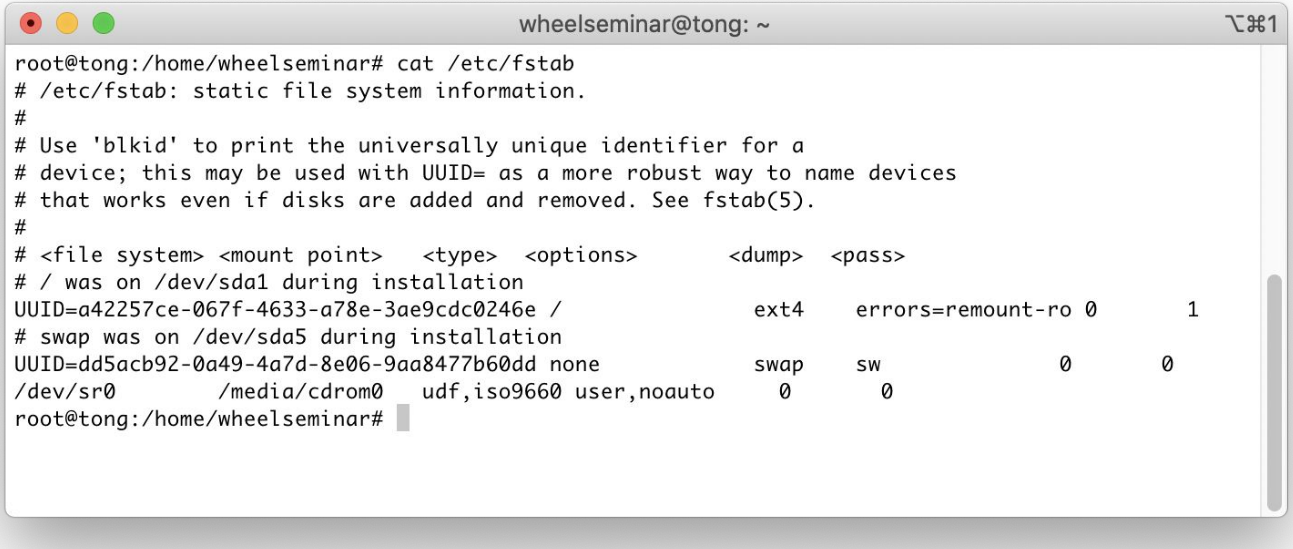
Mount / Unmount device

- Link a partition device and a directory
  => Device can be used as a single directory (mount point)
- Mount automatically: /etc/fstab
- Mount manually: mount

Mount automatically: /etc/fstab

- fstab file: Information of file systems

  <device> <mount point> <file system type> <options> <backup operation> <file system check order>
- Options - auto (default) : mount automatically at bootup



```
root@tong:/home/wheelseminar# 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/sda1 during installation
UUID=a42257ce-067f-4633-a78e-3ae9cdc0246e /               ext4    errors=remount-ro 0       1
# swap was on /dev/sda5 during installation
UUID=dd5acb92-0a49-4a7d-8e06-9aa8477b60dd none            swap    sw              0       0
/dev/sr0        /media/cdrom0   udf,iso9660 user,noauto     0       0
root@tong:/home/wheelseminar# 
```

Mount manually: mount

- mount
  - Current mount information
- mount -t <file system type> <device> <mount point>
- mount -a
  - /etc/fstab auto
- umount <device>
  umount <mount point>

Check and restore file system: fsck

- Check consistency of file system and restore if there is an error
- Always use after unmount
- Fix applied only after reboot

Swap area

Swap area

- Uses part of disk as RAM
- (Traditionally) RAM size * 2
- Allocation of swap area
    - Swap file: Use swap file in file system - Able to allocate while system is running
    - Swap partition: Better performance since disk blocks are contiguous

      => Allocate using fdisk, parted

Allocation of swap area - swap file

```
$ dd if=/dev/zero of=<swap file location> bs=<buffer_size> count=<num_buffers>
```

- If bs=1k, count is file size
- /dev/zero: ASCII NULL (0x00)

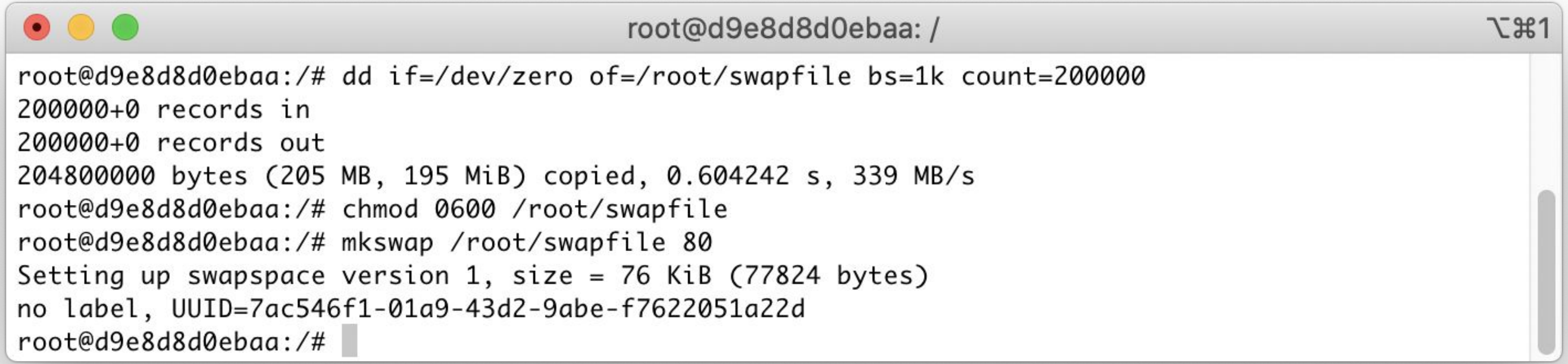
```
$ chmod 600 <swap file location>
```

```
$ mkswap <swap file location> <size in KB>
```

```
$ swapon <swap file location or partition>
```

```
$ swapoff <swap file location or partition>
```

Allocation of swap area - swap file

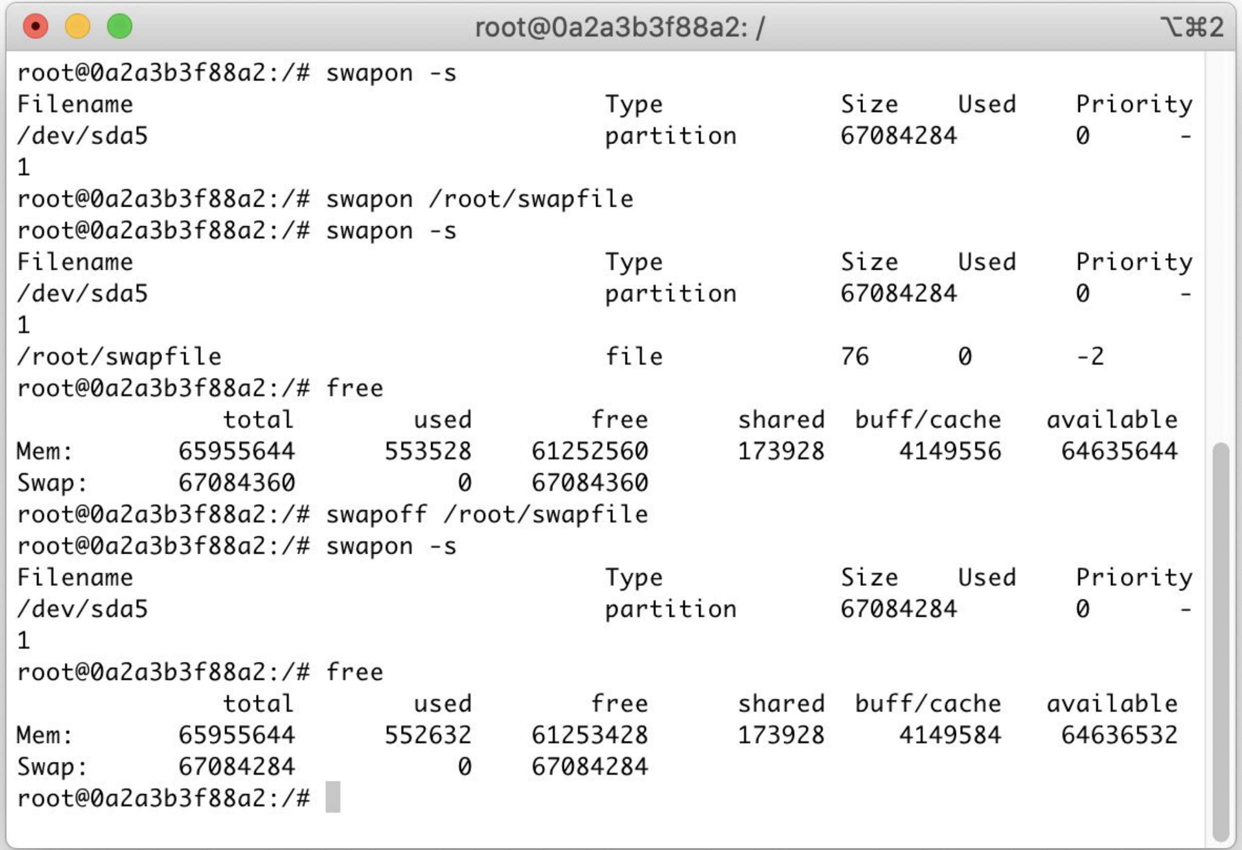


```
root@d9e8d8d0ebaa:/# dd if=/dev/zero of=/root/swapfile bs=1k count=200000
200000+0 records in
200000+0 records out
204800000 bytes (205 MB, 195 MiB) copied, 0.604242 s, 339 MB/s
root@d9e8d8d0ebaa:/# chmod 0600 /root/swapfile
root@d9e8d8d0ebaa:/# mkswap /root/swapfile 80
Setting up swapspace version 1, size = 76 KiB (77824 bytes)
no label, UUID=7ac546f1-01a9-43d2-9abe-f7622051a22d
root@d9e8d8d0ebaa:/#
```

Allocation of swap area - swap file

$ free

$ swapon -s



```
root@0a2a3b3f88a2:/# swapon -s
Filename                                Type            Size    Used    Priority
/dev/sda5                               partition       67084284        0       -
1
root@0a2a3b3f88a2:/# swapon /root/swapfile
root@0a2a3b3f88a2:/# swapon -s
Filename                                Type            Size    Used    Priority
/dev/sda5                               partition       67084284        0       -
1
/root/swapfile                          file            76      0       -2
root@0a2a3b3f88a2:/# free
              total        used        free      shared  buff/cache   available
Mem:       65955644      553528    61252560      173928     4149556    64635644
Swap:      67084360           0    67084360
root@0a2a3b3f88a2:/# swapoff /root/swapfile
root@0a2a3b3f88a2:/# swapon -s
Filename                                Type            Size    Used    Priority
/dev/sda5                               partition       67084284        0       -
1
root@0a2a3b3f88a2:/# free
              total        used        free      shared  buff/cache   available
Mem:       65955644      552632    61253428      173928     4149584    64636532
Swap:      67084284           0    67084284
root@0a2a3b3f88a2:/# 
```

References

- andromeda-20140729-0.pdf
- 2019 Spring CS330 Lecture slides (Youngjin Kwon)