Storage and Filesystem

Michael Tsai
2019/03/10

Storage Hardware

185 TB (Sony)

- Magnetic Tape

- Optical Disks (CD, DVD, Blue-ray)

25/50/100/128 GB

- Hard Drive

12 TB, 3.5 inch (Seagate & WD)

- SSD

30 TB, 2.5 inch (Samsung)

https://www.theverge.com/circuitbreaker/2018/2/20/17031256/worlds-largest-ssd-drive-samsung-30-terabyte-pm1643

HD v.s. SSD

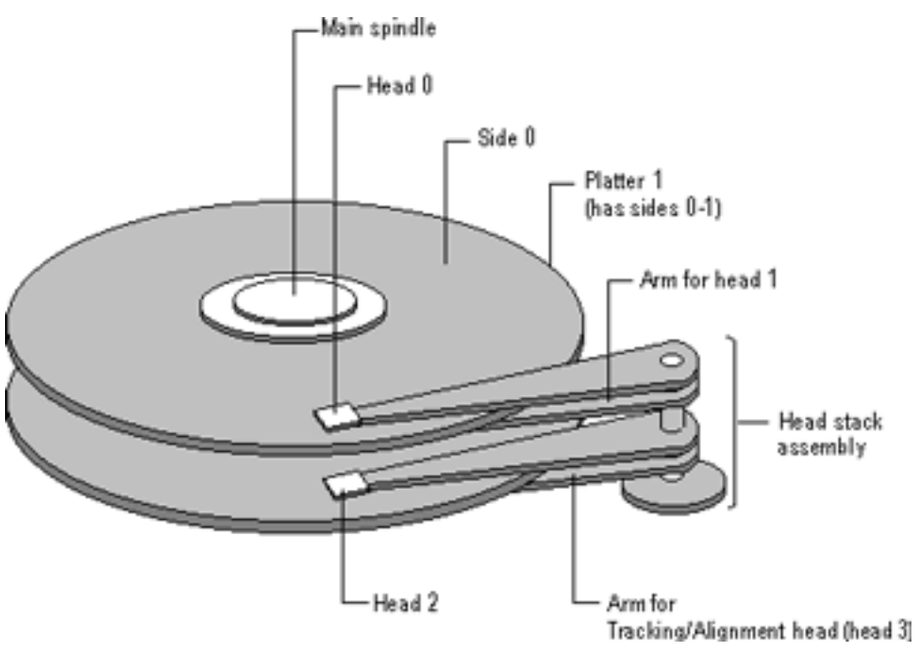
	HD	SSD
Size	Terabytes (Max: 14 TB)	Terabytes (Max: 2 TB)
Random access time	8 ms	0.25 ms
Sequential read	100 MB/s	560 MB/s (M.2 3230 MB/s)
Random read	75-100 IOPS (7,200 rpm) 175-210 IOPS (15,000 rpm) ~ 2 MB/s	100,000 4K IOPS > 30 MB/s
Cost (2019)	~TWD 1.185 / GB (Seagate IronWolf 14 TB)	~TWD 6.5/GB (Intel 760P 2TB M.2 2280)
Limited writes	No	Yes

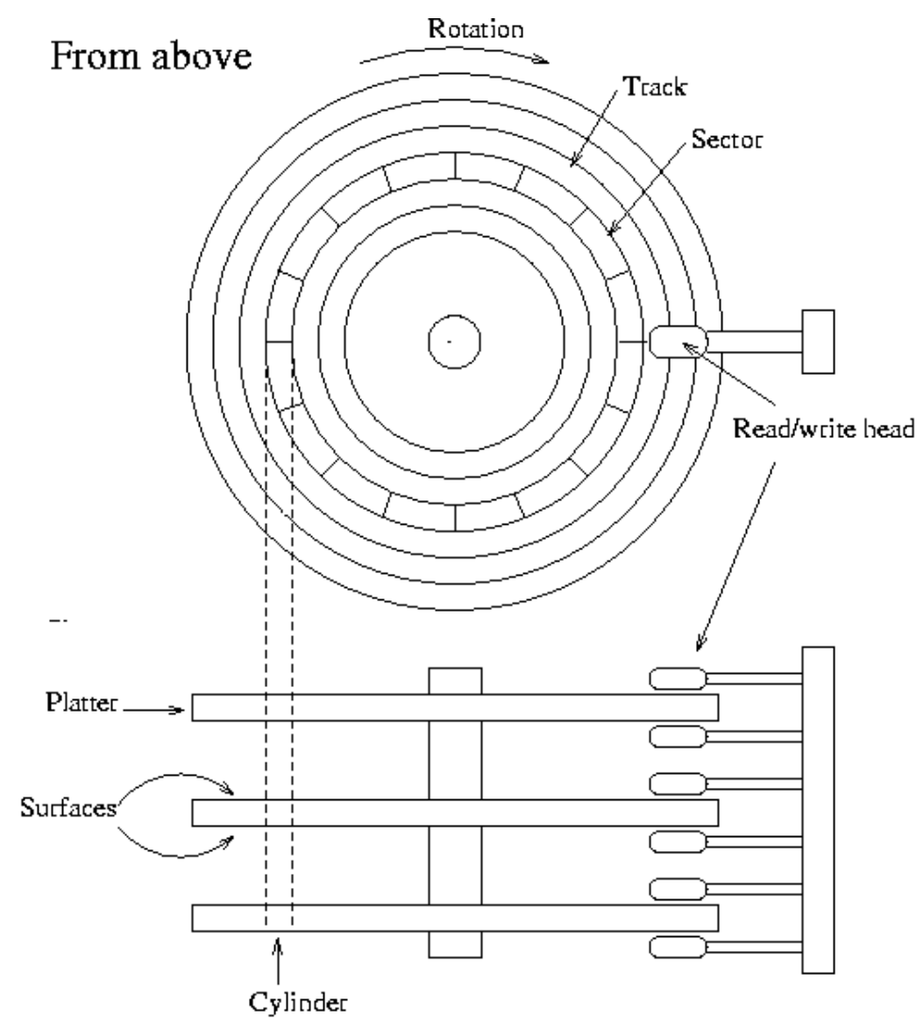
Hard Drive

- History:
  - 60 MB HD = $1,000 USD (~1990) —>14TB HD = ~$612 USD (2019)
  - See here for a graph: (up to 2015) http://www.mkomo.com/cost-per-gigabyte-update
  - Sequential read: 500 kB/s —> 100 MB/s

Hard Drive

- Delay: seek delay and rotational delay
- > 10 MB/s —> < 5 MB/s





HD: other information

- Unit comparison:
  - Disk: Gigabyte = 1,000,000,000 bytes
    Memory: Gigabyte = 2^30 bytes (7% difference)
- Failure statistics: (from 2007 Google Labs study)
  - 2 years (6% average annualized failure rate)
    5 years (less than 75% to survive)
  - Annualized Failure Rate (AFR)=
    estimated probability that a device or component will fail
    during a full year of use
  - **Operating temperature** and **drive activity** are **not** correlated with failure rate
- Read: Backblaze Hard Drive Stats 2018 (Just out in Jan., 2019)
  https://www.backblaze.com/blog/hard-drive-stats-for-2018/
  (AFR for all Backblaze drive models was just 1.25%)

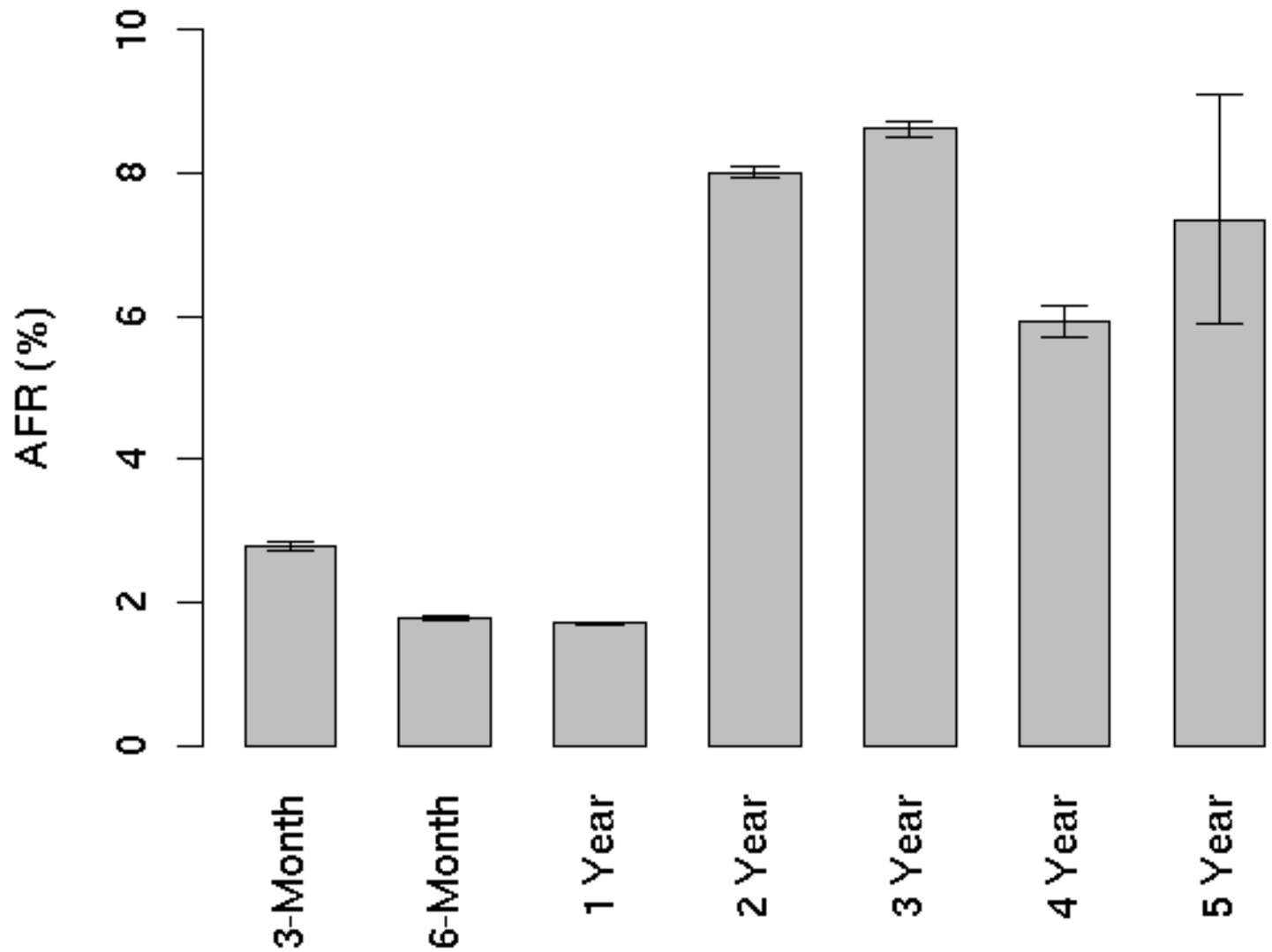


Figure 2: Annualized failure rates broken down by age groups

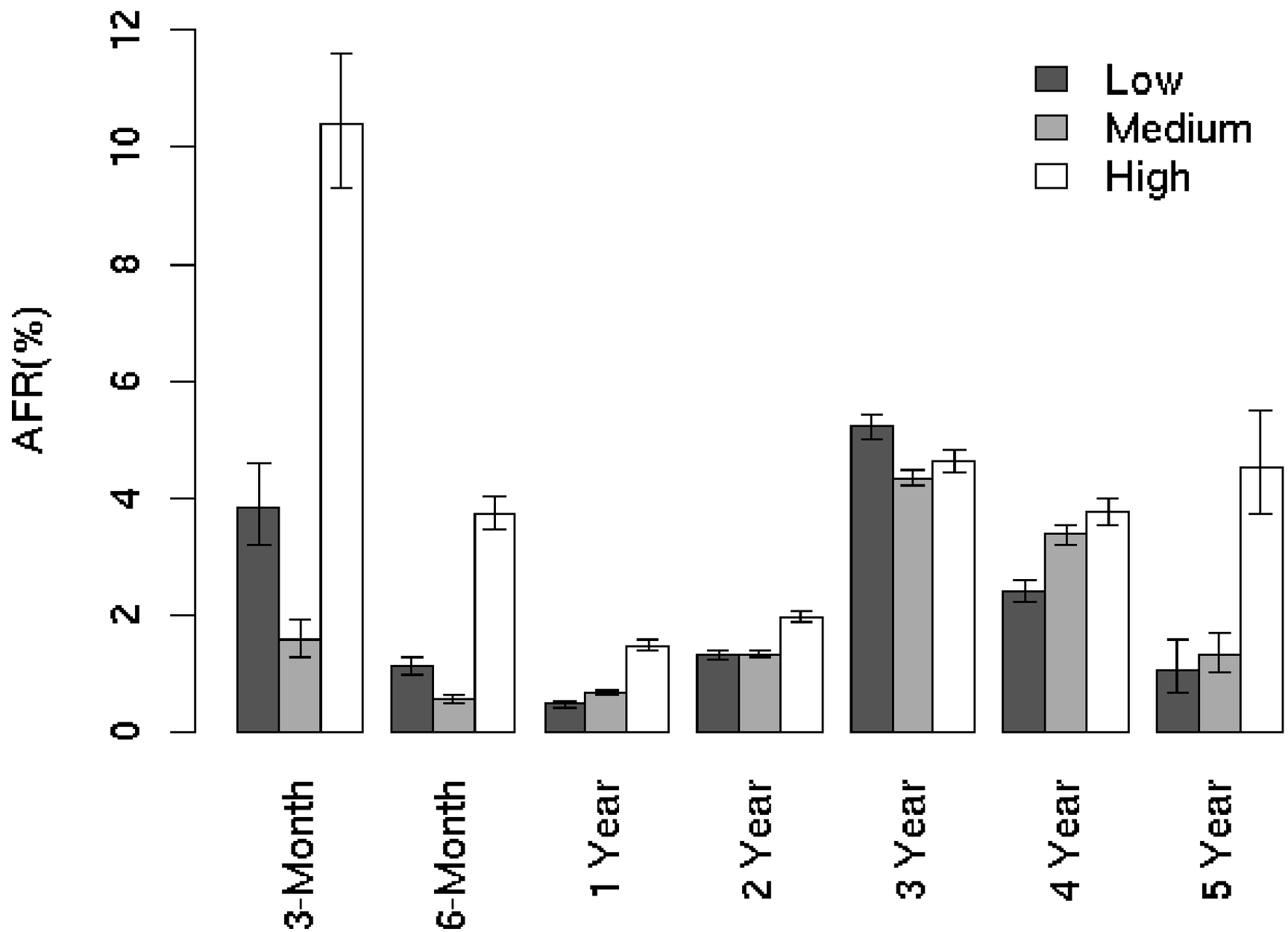


Figure 3: Utilization AFR

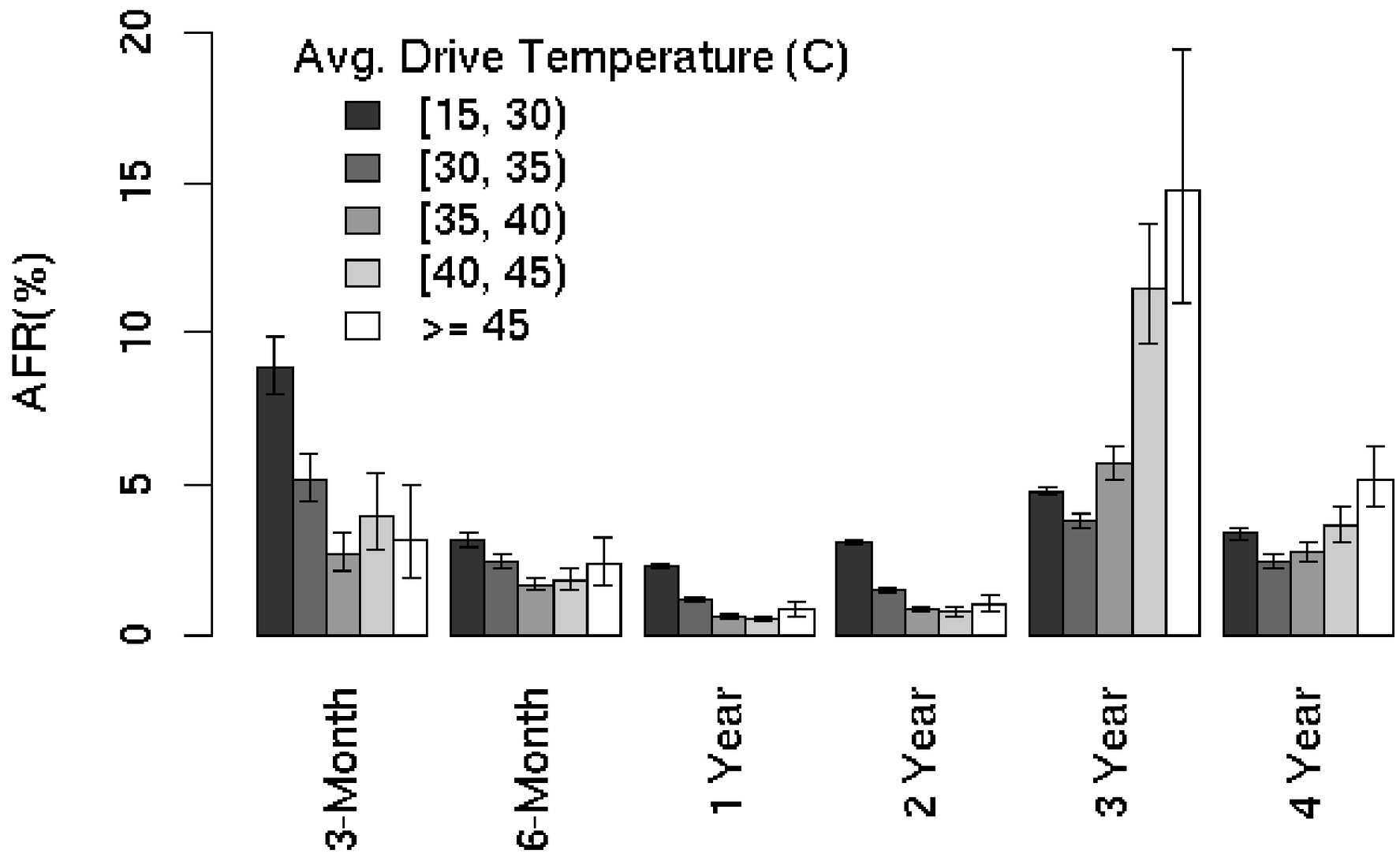


Figure 5: AFR for average drive temperature.

Backup Strategy

- Read: https://www.backblaze.com/blog/the-3-2-1-backup-strategy/

SSD: solid state disks

- Read: How do SSDs work? https://www.extremetech.com/extreme/210492-extremetech-explains-how-do-ssds-work
- Read and write across banks of flash memory cells
- Individually slower than HD, but can use **parallelism**
- Write cycles limitation: 100,000 (typical)
- Firmware spreads the write across all pages
- **Erasure** is required before write (and slower than write)
- Clusters of adjacent pages **HAVE TO** be erased together
- Q: why is a SSD gets slower as we use it more?

2017

MyDigitalSSD BPX (240GB)	MyDigitalSSD BPX (480GB)	Plextor M8Pe (512GB)	Plextor M8Pe (1TB)	Samsung 960 EVO (1TB)
				
REVIEW ›	REVIEW ›	REVIEW ›	REVIEW ›	REVIEW ›
$114.99 Amazon ›	$199.99 Amazon ›	$299.99 Newegg ›	$499.99 Newegg ›	$499.99 Newegg ›

2018

SK hynix SL308 (500GB)	SanDisk Ultra 3D (1TB)	WD Blue 3D (1TB)	Samsung 850 EVO (2TB)	Samsung 850 EVO (4TB)
				
REVIEW ›	REVIEW ›	REVIEW ›	REVIEW ›	REVIEW ›
$155.99 Amazon ›	$249.99 Amazon ›	$259.94 Amazon ›	$745.41 Amazon Marketplace ›	$1,356.98 Amazon ›

SSD: other information

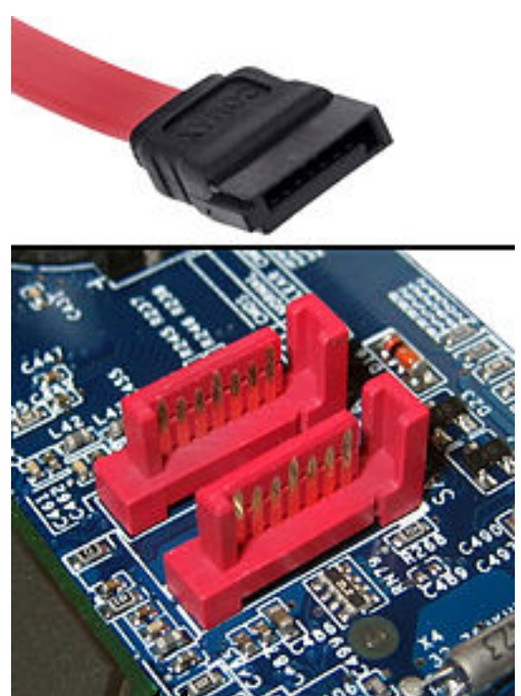
- Filesystem needs to be “SSD-aware”
  - Let it know what blocks are no longer used (erased)
- Alignment: 512 byte blocks (SSD) vs 1~8 KiB (FS)
  - SSD can only read/write 4 KiB pages
  - Need to align the boundaries
- Write cycle limitation: when will it run out?
  - 100 MB/s —> 150 GB SSD for **continuous 4 years.**
    (Some can do 750 TB…)
  - Test: http://techreport.com/review/27436/the-ssd-endurance-experiment-two-freaking-petabytes

Hardware Interface


33,66,100,133 MB/s

- (P)**ATA** (Advanced Technology Attachment) or **IDE** (Integrated Drive Electronics)

- **SATA** (Serial ATA)



1.5,3,6 Gb/s
(150, 300, 600 MB/s)

Other interfaces

- SCSI (5, 20, 40 MB/s)
  SAS (serial attached SCSI)
  (3, 6 Gbps)

- Fiber Channel (1-40 Gb/s)

- USB: 2.0: 480 Mb/s / 3.0: 5 Gb/s / 3.1: 10 Gb/s
  FireWire (IEEE 1394): 400 and 800 MB/s
  Thunderbolt: 1: 10 Gb/s / 2: 20 Gb/s

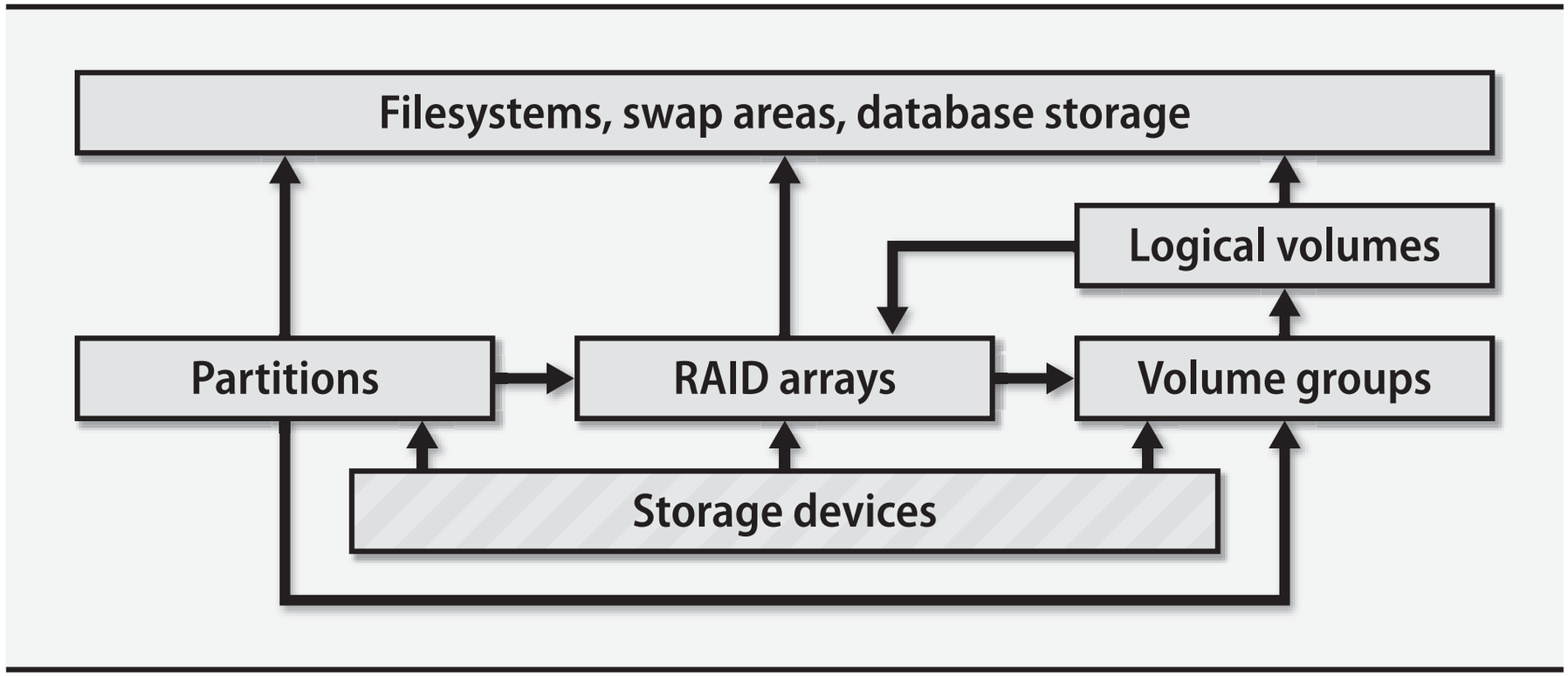
- PCI-express:
  (2,4,8,16 Gb/s)

OCZ RevoDrive 350

Storage management layers

Arrow: “can be built on”

Exhibit B Storage management layers



Storage “pieces”

- Storage device: “disk” -
  random access, block I/O, represented by a device file

- Partition: fixed-size subsection of a storage device
  (古代遺跡: 為了跟windows使用的儲存裝置相容)

- RAID array:
  increase performance, reliability, or both

- Volume groups & logical volumes:
  related to logical volume manager (LVM) /
  aggregation & split

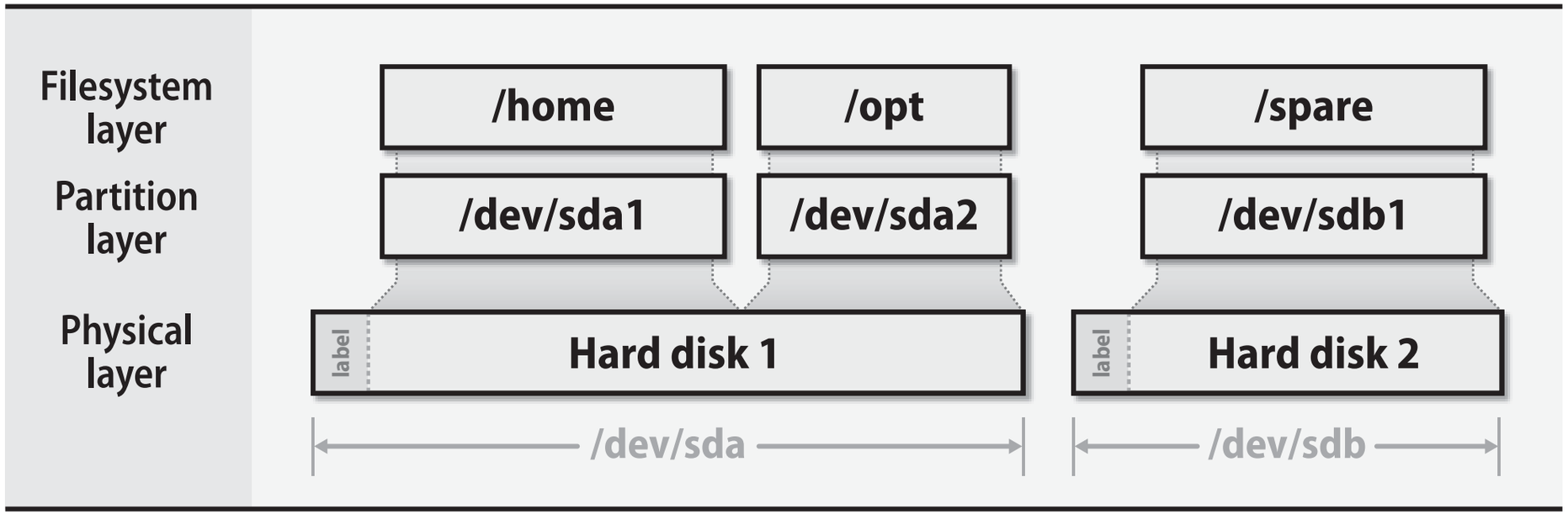
Storage “pieces”

- Filesystem: mediates between
  - blocks presented by a partition, RAID, or logical volume
  - standard filesystem interface expected by programs
    - path: e.g., /var/spool/mail
    - File types, file permissions, etc.
  - how the content of files are stored
  - how the filesystem namespace is represented and searched on disk
  - other “filesystems”: swap and database storage

Table 6.2 Standard directories and their contents

Pathname	OS[a]	Contents
/bin	All	Core operating system commands[b]
/boot	LS	Kernel and files needed to load the kernel
/dev	All	Device entries for disks, printers, pseudo-terminals, etc.
/etc	All	Critical startup and configuration files
/home	All	Default home directories for users
/kernel	S	Kernel components
/lib	All	Libraries, shared libraries, and parts of the C compiler
/media	LS	Mount points for filesystems on removable media
/mnt	LSA	Temporary mount points, mounts for removable media
/opt	All	Optional software packages (not consistently used)
/proc	LSA	Information about all running processes
/root	LS	Home directory of the superuser (often just **/**)
/sbin	All	Commands needed for minimal system operability[c]
/stand	H	Stand-alone utilities, disk formatters, diagnostics, etc.
/tmp	All	Temporary files that may disappear between reboots
/usr	All	Hierarchy of secondary files and commands
/usr/bin	All	Most commands and executable files
/usr/include	All	Header files for compiling C programs
/usr/lib	All	Libraries; also, support files for standard programs
/usr/lib64	L	64-bit libraries on 64-bit Linux distributions
/usr/local	All	Software you write or install; mirrors structure of **/usr**
/usr/sbin	All	Less essential commands for administration and repair
/usr/share	All	Items that might be common to multiple systems
/usr/share/man	All	On-line manual pages
/usr/src	LSA	Source code for nonlocal software (not widely used)
/usr/tmp	All	More temporary space (preserved between reboots)
/var	All	System-specific data and configuration files
/var/adm	All	Varies: logs, setup records, strange administrative bits
/var/log	LSA	Various system log files
/var/spool	All	Spooling directories for printers, mail, etc.
/var/tmp	All	More temporary space (preserved between reboots)

Exhibit C Traditional data disk partitioning scheme (Linux device names)



Partition & logical volume

- Why? Easier to backup / Confine damage
- Tips:
  - Have a backup root device and check if it works
  - Put /tmp on a separate filesystem
    (no backup / size limit)
  - Separate /var: log in /var easily fill up /
  - Splitting swap on multiple physical disks / add more swap when adding memory

Partition: other information

- When is it used nowadays?
  - share a disk with windows
  - specify location on the disk
    (outer cylinder is faster by 30%!)
  - create partitions of identical size (for RAID)

Partition: other information

- MBR (windows-style) partition table
  - primary and extended partitions
  - OS is installed in primary partition
  - one partition is marked as “active” and boot loader looks for that partition
  - does not support disk > 2 TB
  - Max # of partitions: 4

Partition: other information

- GPT: GUID partition table
  - support disk > 2TB
  - Windows Vista and versions afterwards support GPT disks for data, but need EFI firmware (new computers) to boot
- Tools: gparted (GUI), parted (command-line), fdisk (does not support GPT)

Logical volume management

- Volume groups (VG): storage devices put into groups
- Logical volumes (LV): assign blocks in VG to LV
- Then LV, as a block device, is used by filesystem
- Powerful features:
    - Move LV among physical devices
    - Grow and shrink LV on the fly
    - Snapshot
    - Replace on-line drives
    - Mirroring / stripping

Typical sequence

- sudo pvcreate /dev/sdb
  (define /dev/sdb to be used)
  (you can also use /dev/sdb2, for example, to use just partition 2 on sdb)

- sudo vgcreate hsinmu /dev/sdb
  (put /dev/sdb into a new VG called hsinmu)

- sudo lvcreate -L 8G -n test_lv hsinmu
  (create a 8G LV in hsinmu called test_lv)

- sudo mkfs -t ext4 /dev/hsinmu/test_lv
  (format the new LV as a ext4 filesystem)

- sudo mkdir /mnt/test_lv

- sudo mount /dev/hsinmu/test_lv /mnt/test_lv (掛載)

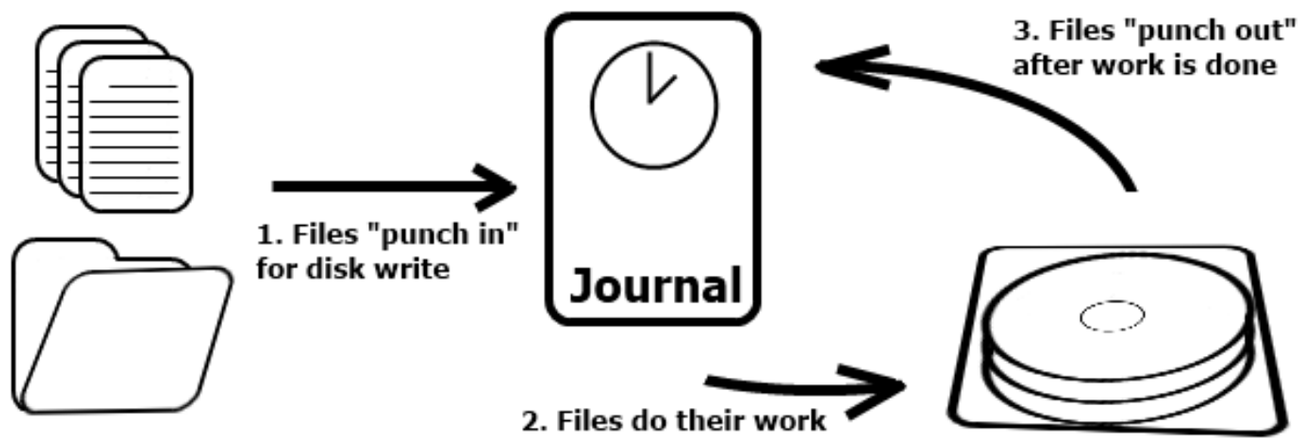
- df -h /mnt/test_lv (show information about a mount point)

Additional reading

- Not covered today:
  - How to do volume snapshots
    (create a copy-on-write duplicate of LV)
    lvcreate -L 8G -s -n snap hsinmu/test_lv
  - Resize the filesystem (lvresize, lvextend)

Filesystem

Journaling filesystem



- Popular filesystems on Linux:
  - Ext 2/3/4 (journaling after 3, better support for SSD in 4)
  - BtrFS (Oracle, better performance - B-Tree FS for server file system, some ReiserFS pros added)
  - ReiserFS / XFS / ZFS