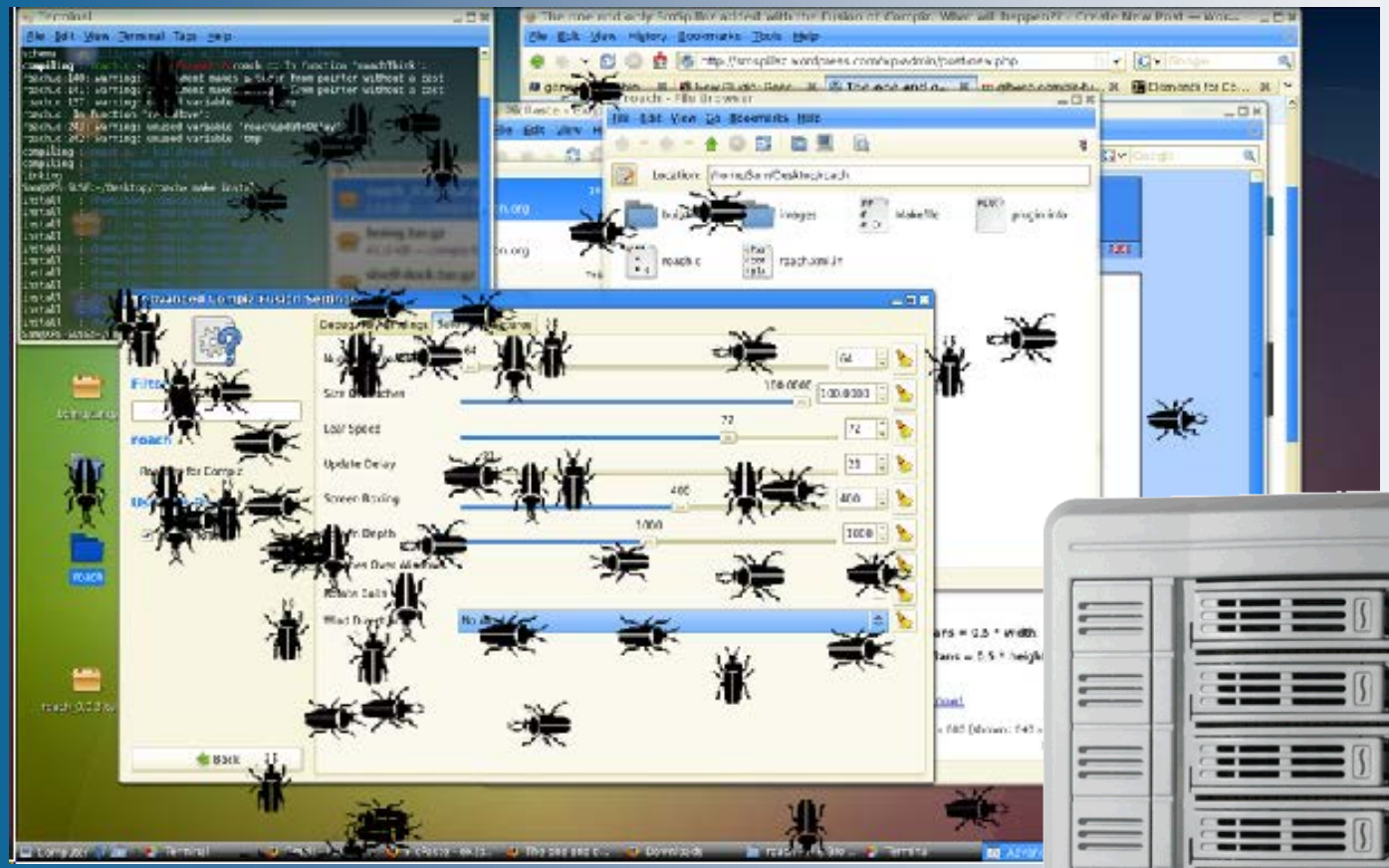
COMP 175

System Administration and Security

RAID, Booting

001000001 001000001 01010010 01010010 01010010



xroaches

RAID

RAID

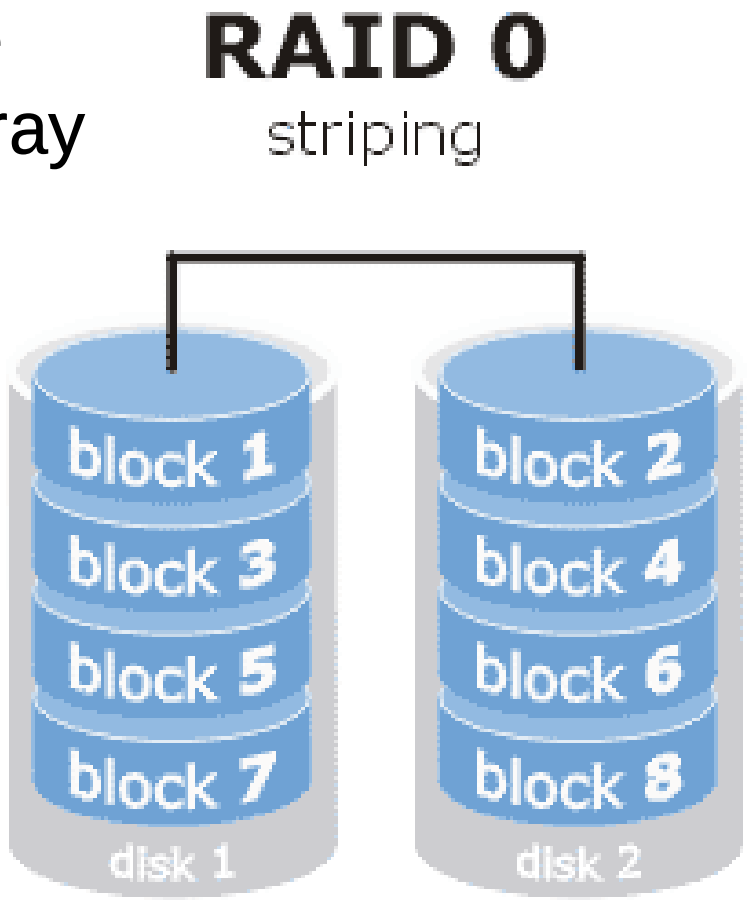
RAID (**R**edundant **A**rray of **I**nexpensive **D**isks)

- Technology used to increase the performance and/or reliability of data storage
- There are three ways to create RAID:

1. Software-RAID: created by software drivers
2. Hardware-RAID: A special controller used to build RAID. Hardware RAID is generally faster, and does not place load on the CPU
3. FakeRAID: RAID hardware is expensive, many motherboard manufacturers use multi-channel controllers with special BIOS features for RAID. Not necessarily faster than true software RAID.

RAID 0: Striping

- Data split into blocks that are written across all drives in array
- Using multiple disks at same time increases I/O performance.
- Use of multiple controllers, ideally one controller per disk, would further increase performance.

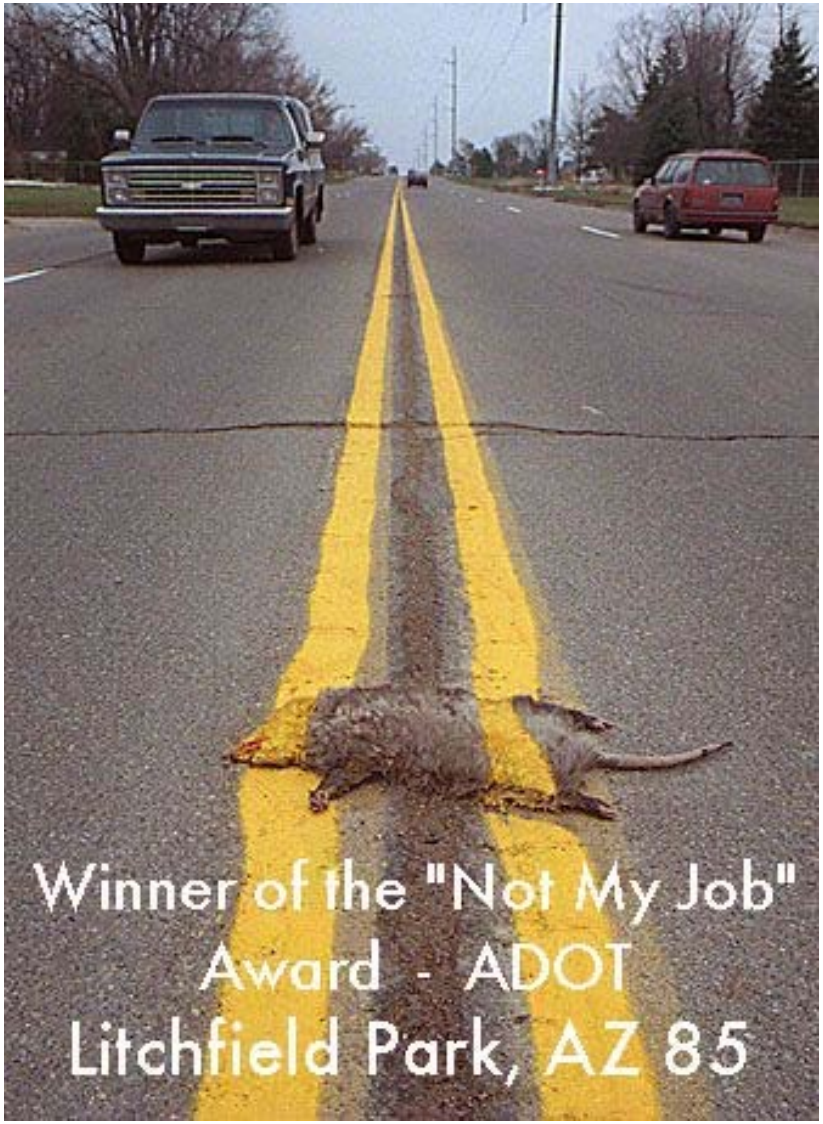


Striping

- Data Striping: the technique of segmenting logically sequential data, such as a file, across different physical or logical storage devices

- Has a single ‘p’
- strīp not strĭp
- Not ecdysiast

  ec·dys·i·ast ekˈdēzēəst



RAID 0

Advantages

- Increased read and write performance operations
- There is no overhead caused by parity controls
- All storage capacity used, no disk overhead
- The technology is easy to implement

Disadvantages

- **Not fault-tolerant**. If one disk fails, all RAID 0 array data lost. Not for mission-critical systems

Ideal use

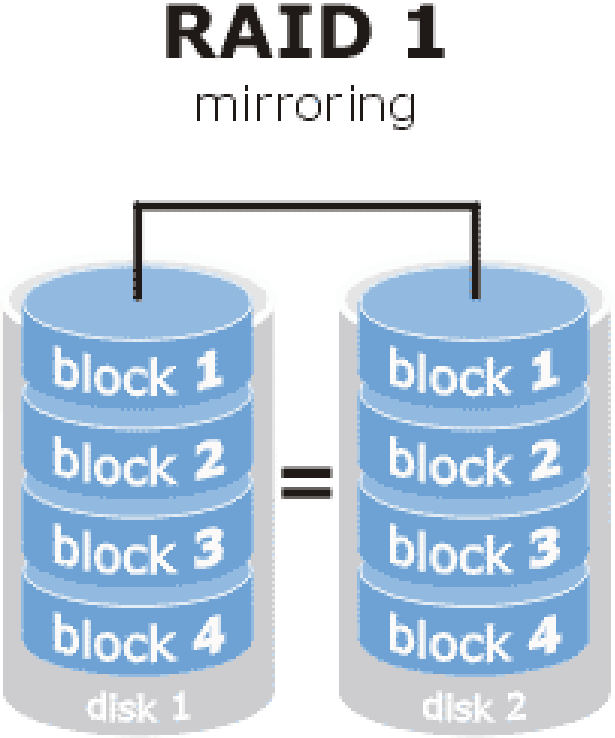
- High speed transient data applications

Speed

Risk

RAID 1: Mirroring

- Data stored twice by writing to both data disk(s) and a mirror disk(s)
- If a disk fails, the controller uses either data drive or mirror drive for data recovery, and continues operation
- Need at least 2 disks for a RAID 1 array



RAID 1

Advantages

- RAID 1 R/W speed comparable to a single disk's
- No rebuild if disk fails, just copy data to new disk
- RAID 1 is simple technology

Disadvantages

- Storage capacity is only half of the total array
- Software RAID 1 may not allow hot swapping *

Ideal use

- RAID-1 is ideal for mission critical storage

* Replacing components without powering down

  Cisco cards: Online Insertion and Removal

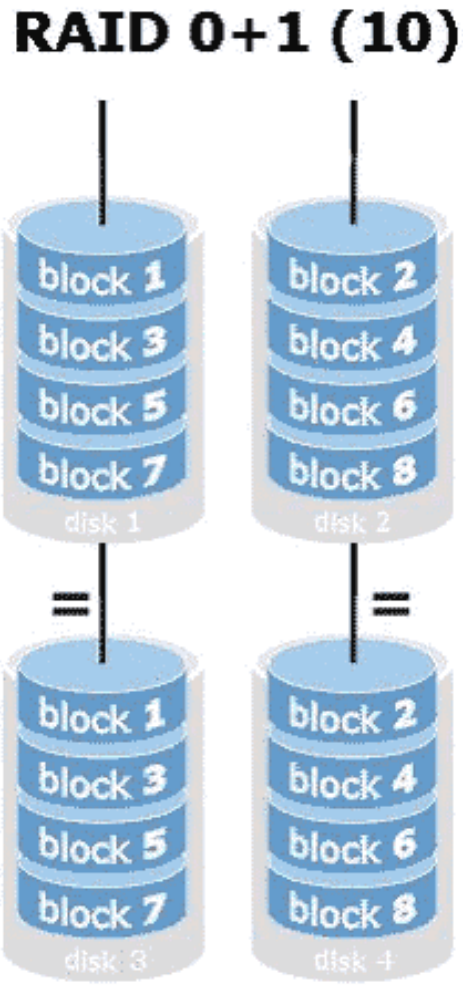
RAID 10

Combine:

- RAID 0 high performance
- RAID 1 data redundancy

- A RAID 10 system
- Same disadvantages

- Does this scale well?

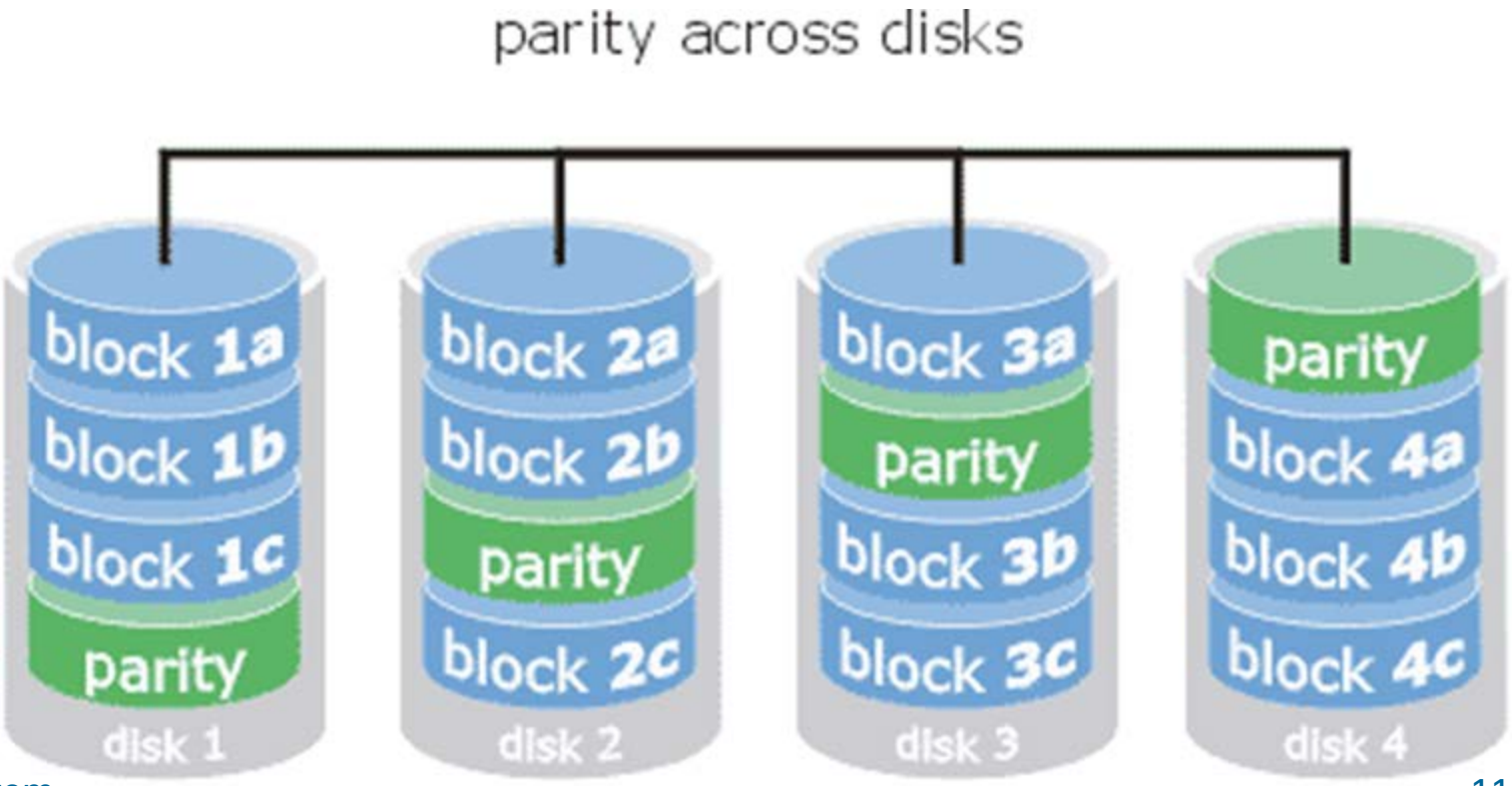


RAID 5

- The most common secure RAID level
- Datablocks are subdivided (striped) and written across multiple drives
- Parity information is spread across all the drives
- Need at least 3 disks for a RAID 5 array
- Array can withstand a single disk failure without losing data or access to data
- Can be achieved in software, but extra cache memory used on hardware controllers improves write performance

RAID 5

- Parity data is used to achieve redundancy
- If a drive fails the remaining data on other drives combined with parity data (using the Boolean XOR function) can reconstruct the missing data



Parity Time

To calculate parity data for the two drives, an XOR is performed on their data:

The resulting parity data then stored on Drive 3

Drive 1: 01101101
Drive 2: 11010100

01101101
XOR 11010100
Drive 3: 10111001

Parity Time

To calculate parity data for the two drives, an XOR is performed on their data:

The resulting parity data then stored on Drive 3

Drive 1: 01101101

Drive 2: 11010100

```
  01101101
XOR 11010100
```

Drive 3: 10111001

Should Drive 2 fail – data from Drive 1 and Drive 3 can recreate Drive 2

```
  10111001 Drive 3
XOR 01101101 Drive 1
  11010100
```

Parity Time

Drive 1: 01101101

Drive 2: 11010100

01101101

XOR 11010100

Drive 3: 10111001

Should Drive 2 fail – data from Drive 1 and Drive 3 can recreate Drive 2

10111001 Drive 3

XOR 01101101 Drive 1

11010100

A Parity Animal!

RAID 5

Advantages

- Read data transactions are very fast , writing data is slower due to parity calculations
- Can run with failed drive (for a while)

Disadvantages

- Disk failures have an effect on throughput
- Is a complex technology

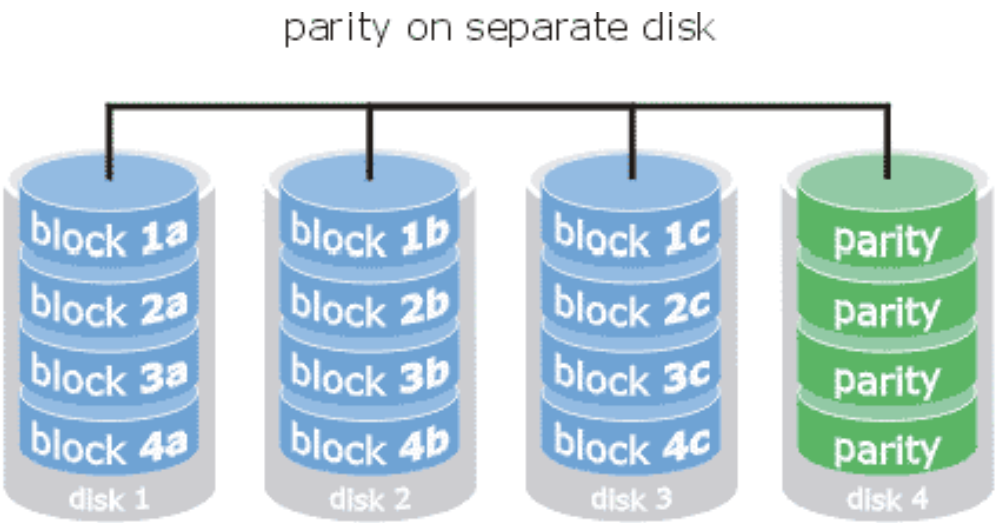
Ideal use

- File and application servers

RAID 3

RAID 3:

- Data blocks are subdivided (striped)
- Written in parallel on two or more drives
- Additional drive stores parity information
- Need at least 3 disks for a RAID 3 array
- Slow for random small I/O transactions



RAID

- RAID-systems can use several interfaces, including SCSI, IDE, SATA or FC (fibre channel.)
- Systems that use SATA disks internally may use a FireWire or SCSI-interface for the host system.
- Sometimes disks in a RAID system are defined as JBOD, ‘Just a Bunch Of Disks’. This means that those disks do not use a specific RAID level and are used as if they were stand-alone disks. This is often done for disks that contain swap files or spooling data.
- RAID is no substitute for Back-Up!

Logical Volume Manager

- LVM can install a bootable system with a root filesystem on a logical volume
- LVM's manage large hard disk farms. Add disks, replace disks, and move data without disrupting service (hot swapping)
- Resize partitions as needed
- Making backups by taking "snapshots“
- Create single logical volumes of multiple physical volumes or entire hard disks (somewhat similar to RAID 0, but more similar to JBOD), allowing for dynamic volume resizing

Direct Attached Storage

- DAS: a JBOD storage array
- Attached to servers HostBusAdapter
- Protocols:
  - SCSI
  - SATA
  - SAS (Serial Attached SCSI)
  - FiberChannel (FCAL)
- Expandable
- Embedded RAID controllers offload RAID processing from HBA
- A virtual hard drive - LUN

Network Attached Storage

- NAS: Network attached file server appliance
- File-level (compared to block level)
- Typically offer RAID
- Often a stripped-down Linux OS
- Protocols: SMB/CIFS, AFP, NFS
- Open source: TurnKey (Ubuntu-based), FreeNAS
- SOHO, Home market growing

Storage Attached Network

- SAN: a separate dedicated network providing access to consolidated block level data storage
- Consolidated DAS from multiple servers
- Increased storage capacity utilization
- Boot from SAN – faster server swap out
- Disaster recovery via distributed storage
- Uses Fiber Channel fabric (w/ FC switches)
- Storage Virtualization
  - Abstracting logical storage from physical storage

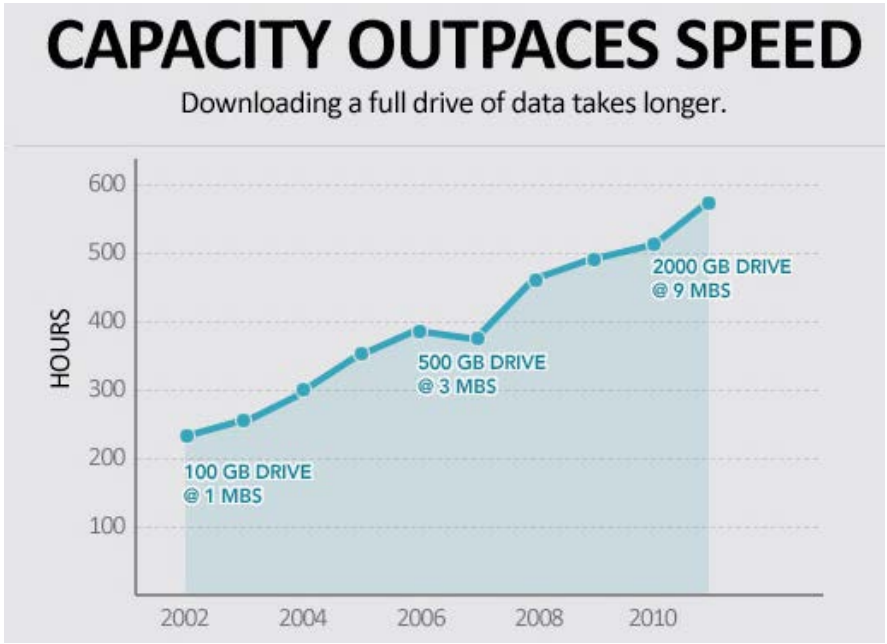
Bytes

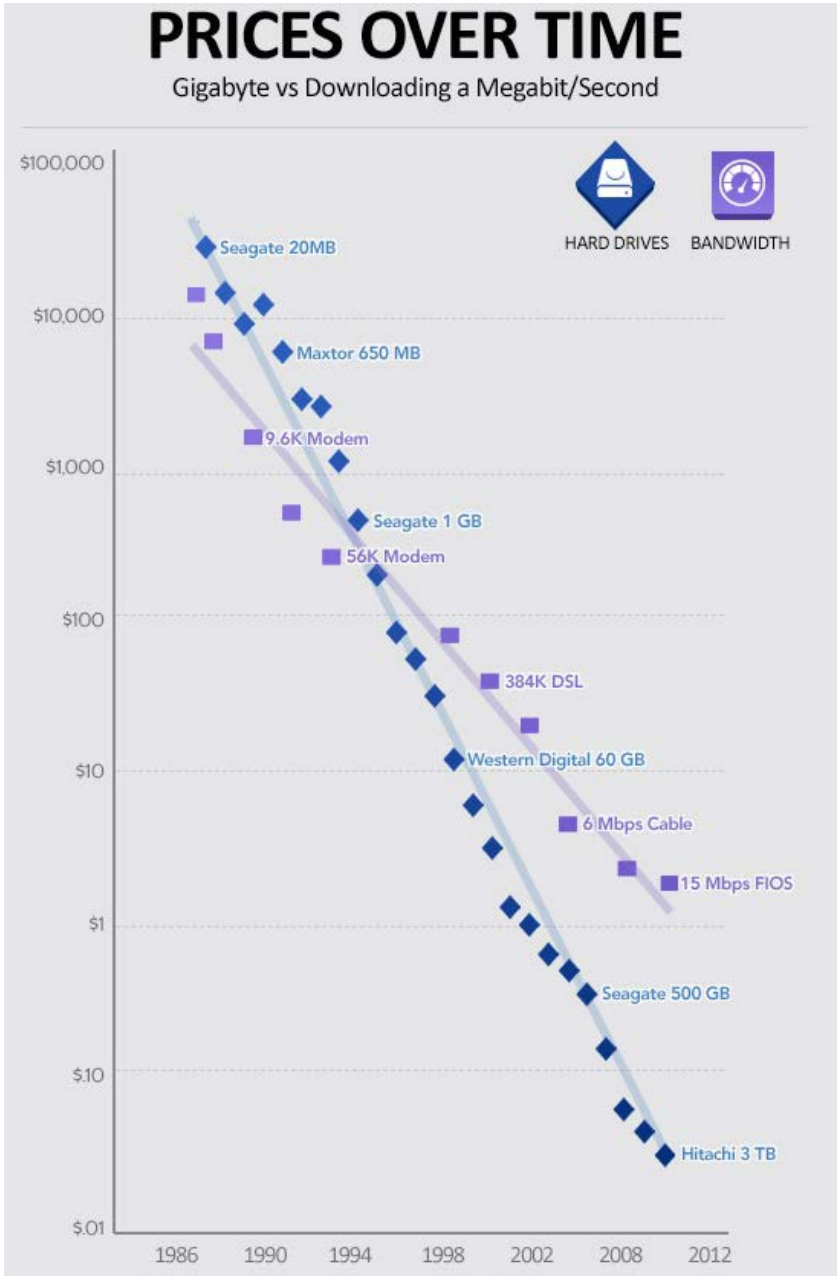
- 1 bit 4 bits = nibble byte = 8 bits
- 1024 Bytes Kilobyte KB
- 1024 KB Megabyte MB
- 1024 MB Gigabyte GB
- 1024 GB Terabyte TB
- 1024 TB Petabyte PB
- 1024 PB Exabyte EB
- 1024 EB Zettabyte ZB
- 1024 ZB Yottabyte YB

Observations

- Storage is cheap
  - Hitachi 3TB 5400 RPM HD $120 (2012)
  - Hitachi 3TB 7200 RPM SATA-6 $120 (2013)
  - Hitachi 3TB 7200 RPM SATA-6 $ 85 (2015)
  - Hitachi 3TB 7200 RPM SATA-6 $ 59 (2018)
  - Hitachi 3TB 7200 RPM SATA-6 $ 37 (2022)
- Efficiency is expensive
  - Eliminating duplication - deduplication
  - File compression
  - Disk compression
- Information life-cycle
- Storage (media) life-span
  - Legacy hardware available? whoops

Storage


CAPACITY OUTPACES SPEED
Downloading a full drive of data takes longer.
HOURS
600
500
400
300
200
100
2002
2004
2006
2008
2010
100 GB DRIVE
@ 1 MBS
500 GB DRIVE
@ 3 MBS
2000 GB DRIVE
@ 9 MBS


PRICES OVER TIME
Gigabyte vs Downloading a Megabit/Second
HARD DRIVES
BANDWIDTH
$100,000
$10,000
$1,000
$100
$10
$1
$.10
$.01
1986
1990
1994
1998
2002
2008
2012
Seagate 20MB
Maxtor 650 MB
9.6K Modem
Seagate 1 GB
56K Modem
384K DSL
Western Digital 60 GB
6 Mbps Cable
15 Mbps FIOS
Seagate 500 GB
Hitachi 3 TB

Remember fsck?

- In ‘89 a company porting UN*X to their hardware often had crashes leaving file system in disarray
- The fsck prompts ‘do you want to clear the file’, etc., got to be a hassle. One of the programmers added a "-y" option to fsck that would print out yes, and automatically clear the file in question and continue.
- This cut reboot times down dramatically
- Until the first: Directory "/“ corrupted, do you wish to remove? YES Directory "/" removed.
- "-y" option removed shortly afterwards

Storage Interfaces

MFM Modified Frequency Modulation

- First 5.25" hard drive Seagate ST-506
- 1980 5Mbit/s $1500 5MB RLL

SCSI Small Computer System Interface

- 1986 40Mbit/s Fast Wide Ultra
- 1994 FC-AL Fiber Channel Arbitrated Loop

IDE Integrated Drive Electronics

- 1986 24Mbit/s
- AT Attachment Interface 1986 41.6Mbit/s
- Renamed PATA Parallel ATA in 2003
- Ultra DMA ATA 133 1,064Mbit/s

SATA Serial ATA

- SATA 1.0 2003 1.5Gbit/s
- SATA 2.0 2010 3Gbit/s
- SATA 3.0 2011 6Gbit/s revision 3.5 in 2020

Interfaces

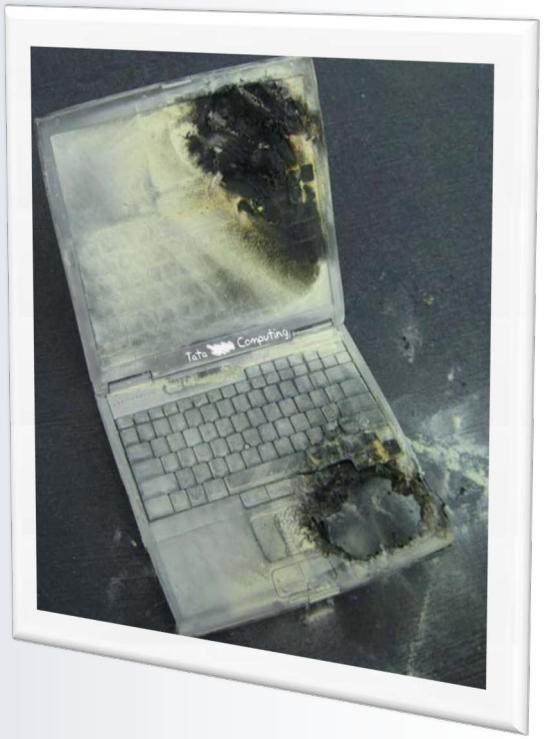
- MFM - Modified Frequency Modulation
- IDE/ATA – Integrated Drive Electronics
- SATA 3, 6 – Serial ATA
- Firewire
- FCP – Fiber Channel Protocol
  - (FC-AL) - Fibre Channel Arbitrated Loop
- UAS – USB Attached SCSI
- Parallel SCSI (clock skew issues w/cabling, termination)
- SCSI, SCSI 2, SCSI 3, SCSI Fast, SCSI Wide, iSCSI

Storage

Mauritian Sunset, 2006

Back side of Mauritian Sunset
Sandy Smith: New York based visual artist

Green/Blue Horizontal, 2005



Hardcore Linux Filesystems

Hardcore Linux Filesystems

COMP 175 *is* an upper division Computer Sciences course

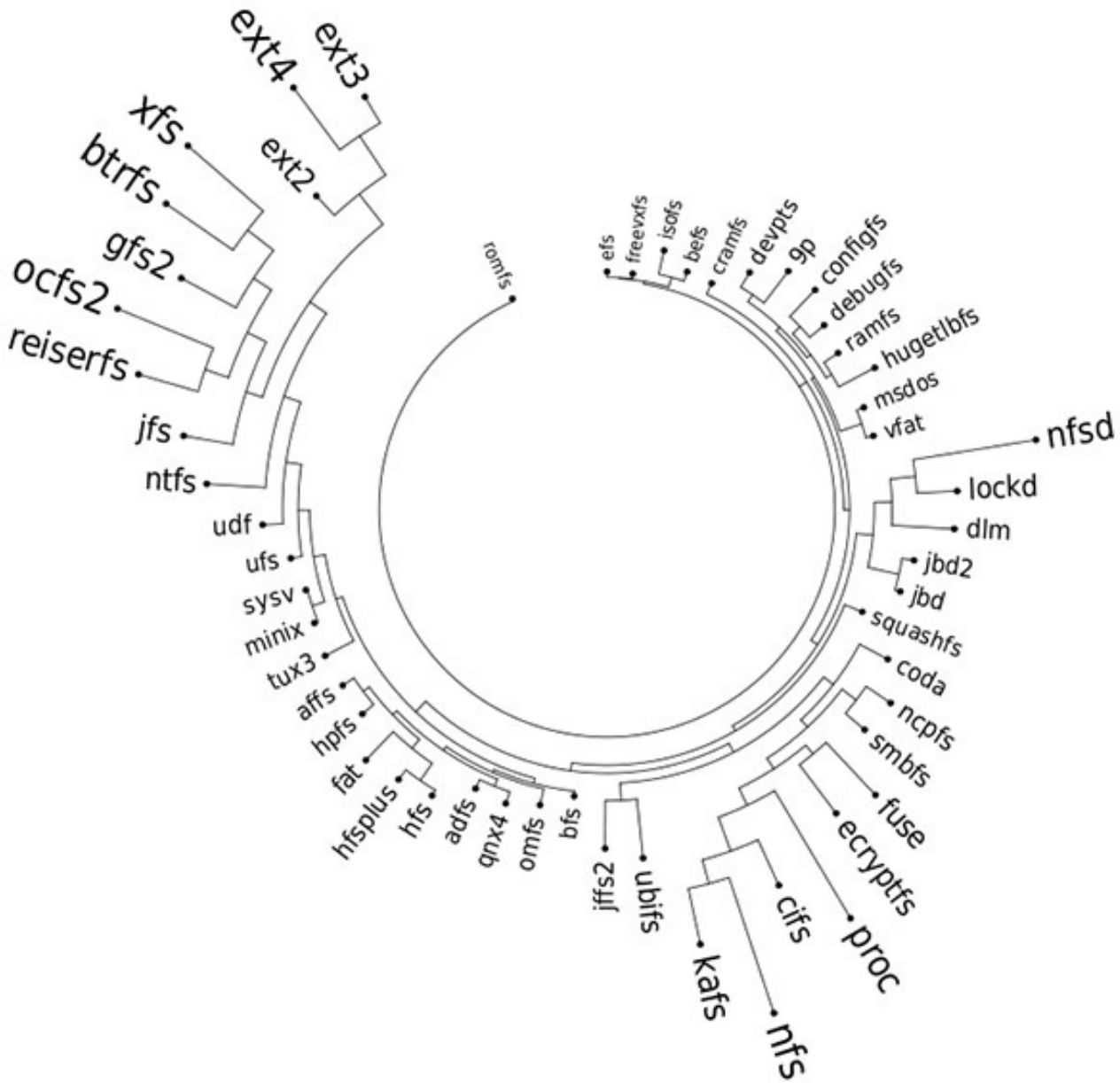
- Harvard: Storage and File System Design Research
- Berkeley: Advanced Topics in Computer Systems: Persistent Storage

thus:

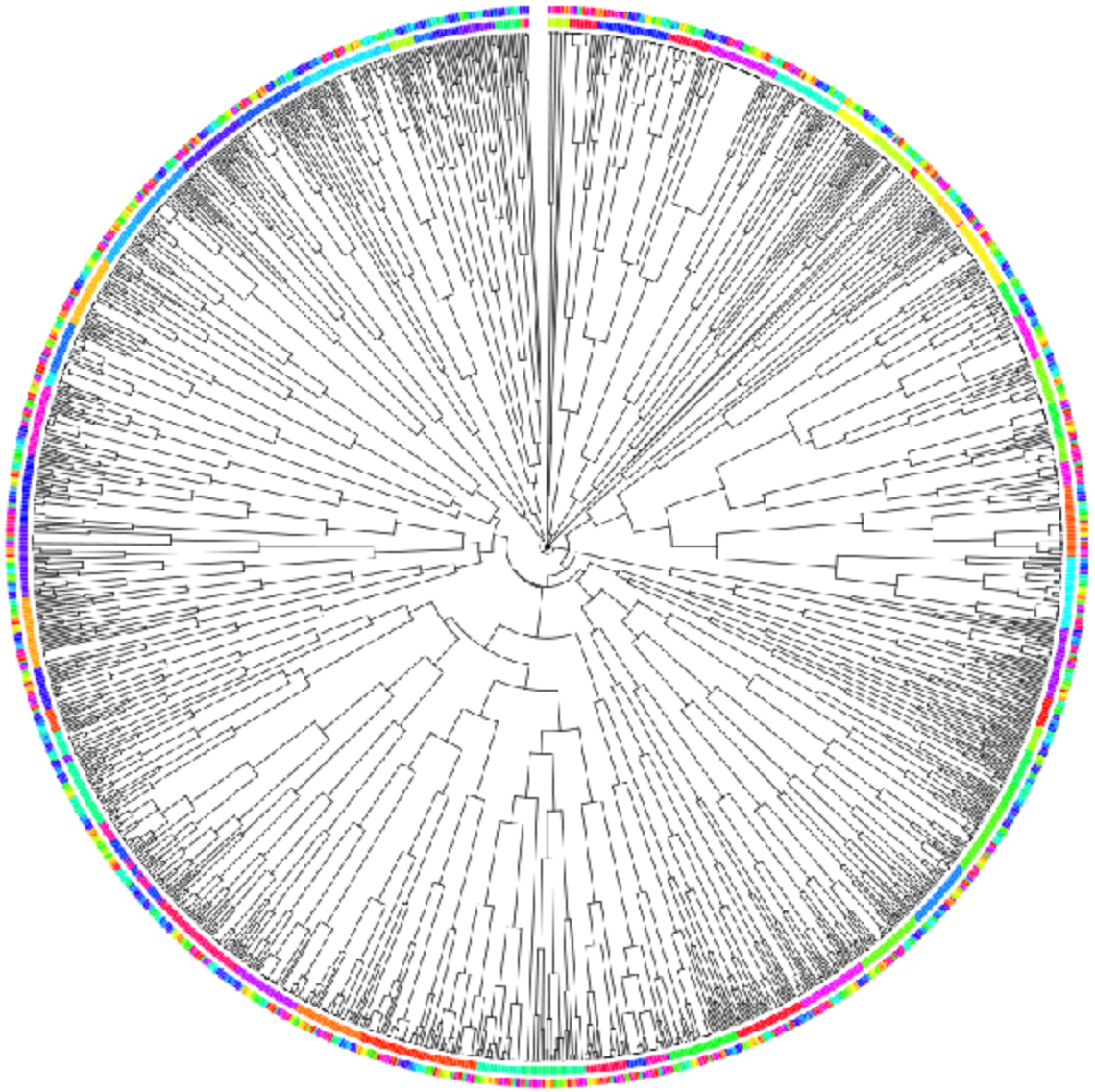
A Visual Expedition Inside the Linux Filesystems

http://cs.jhu.edu/~razvanm/fs-expedition/

Visual Linux File Systems

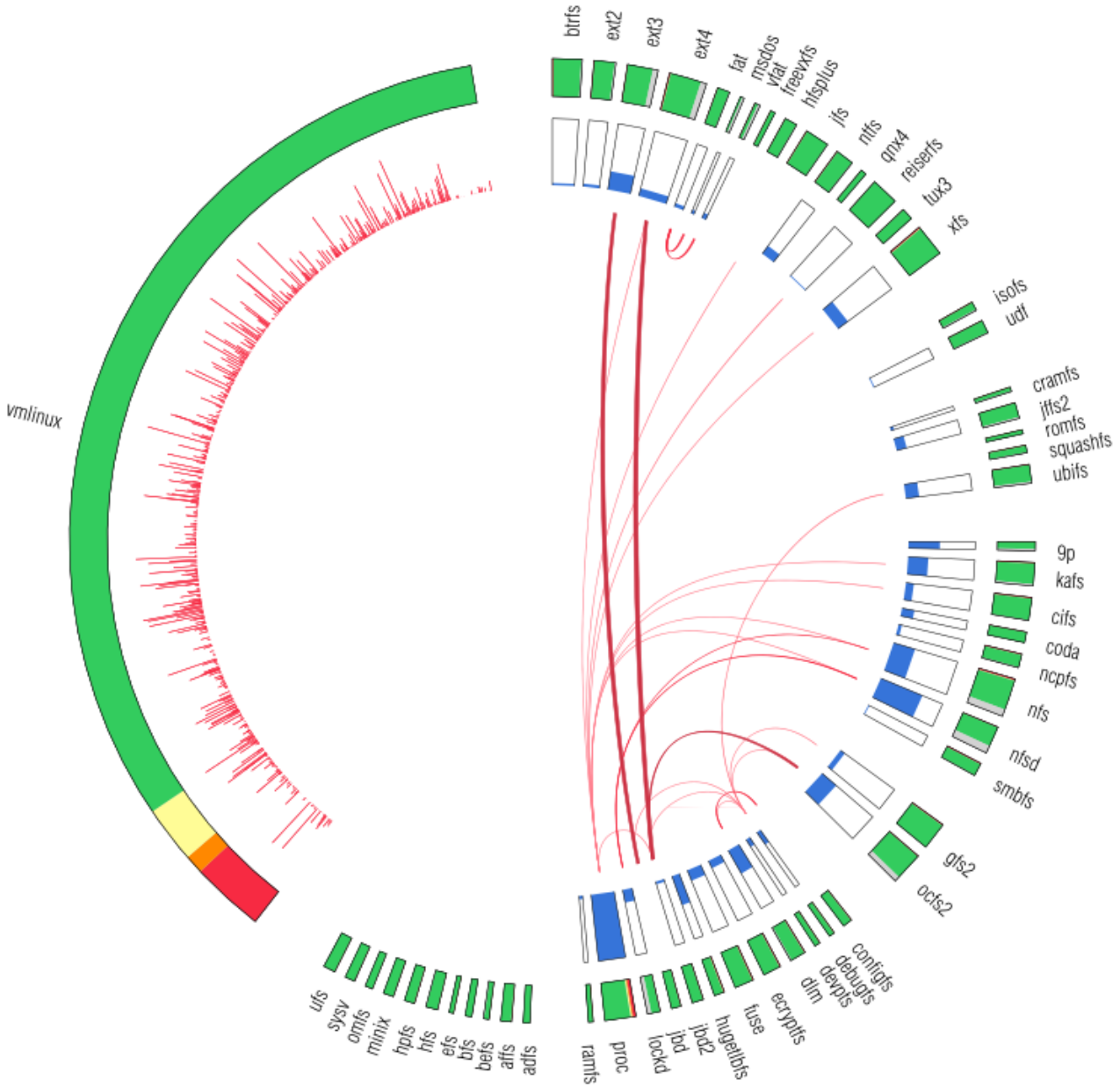


Visual Linux File Systems

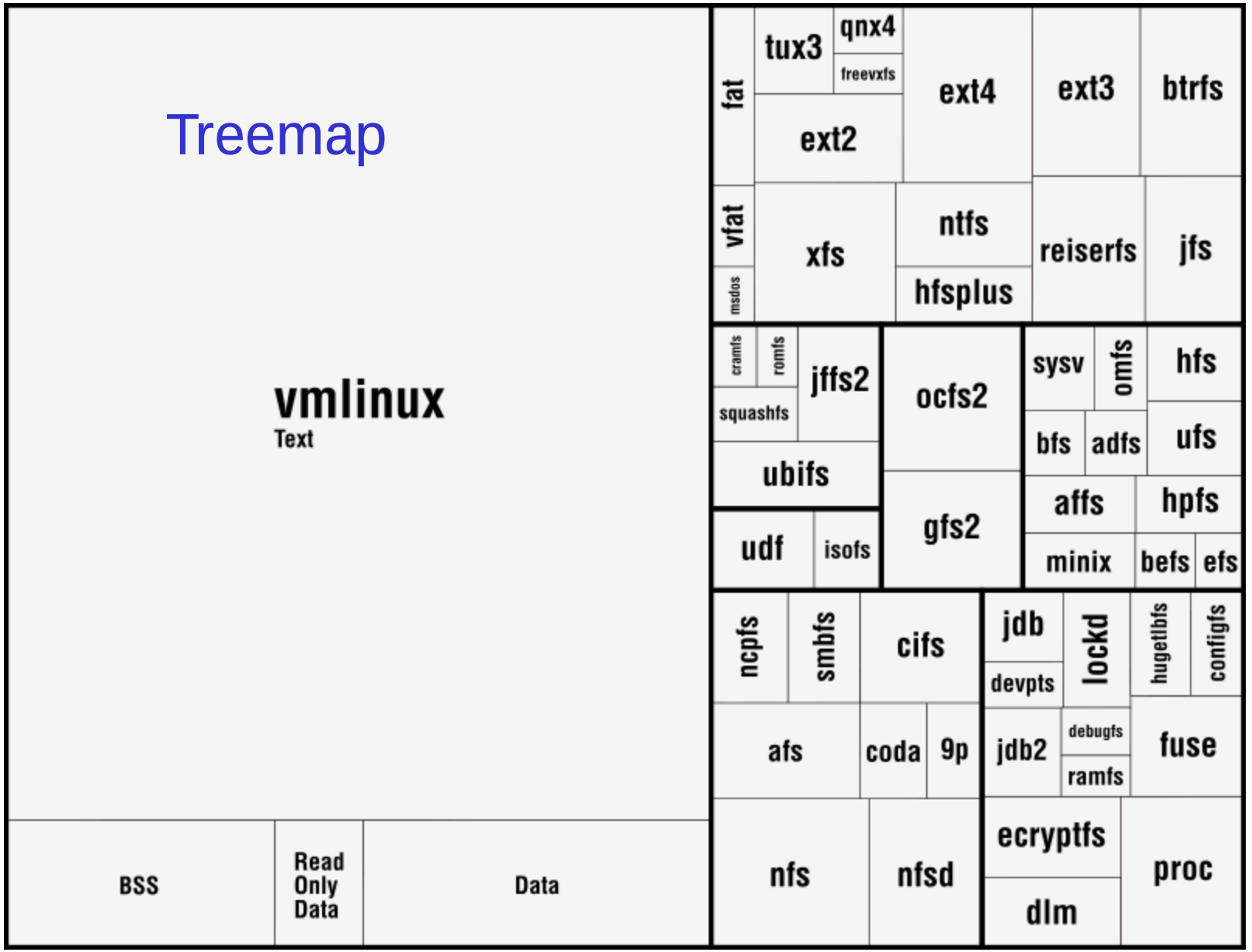


Circular dendrogram of the clustering using Canberra distance and complete linkage.

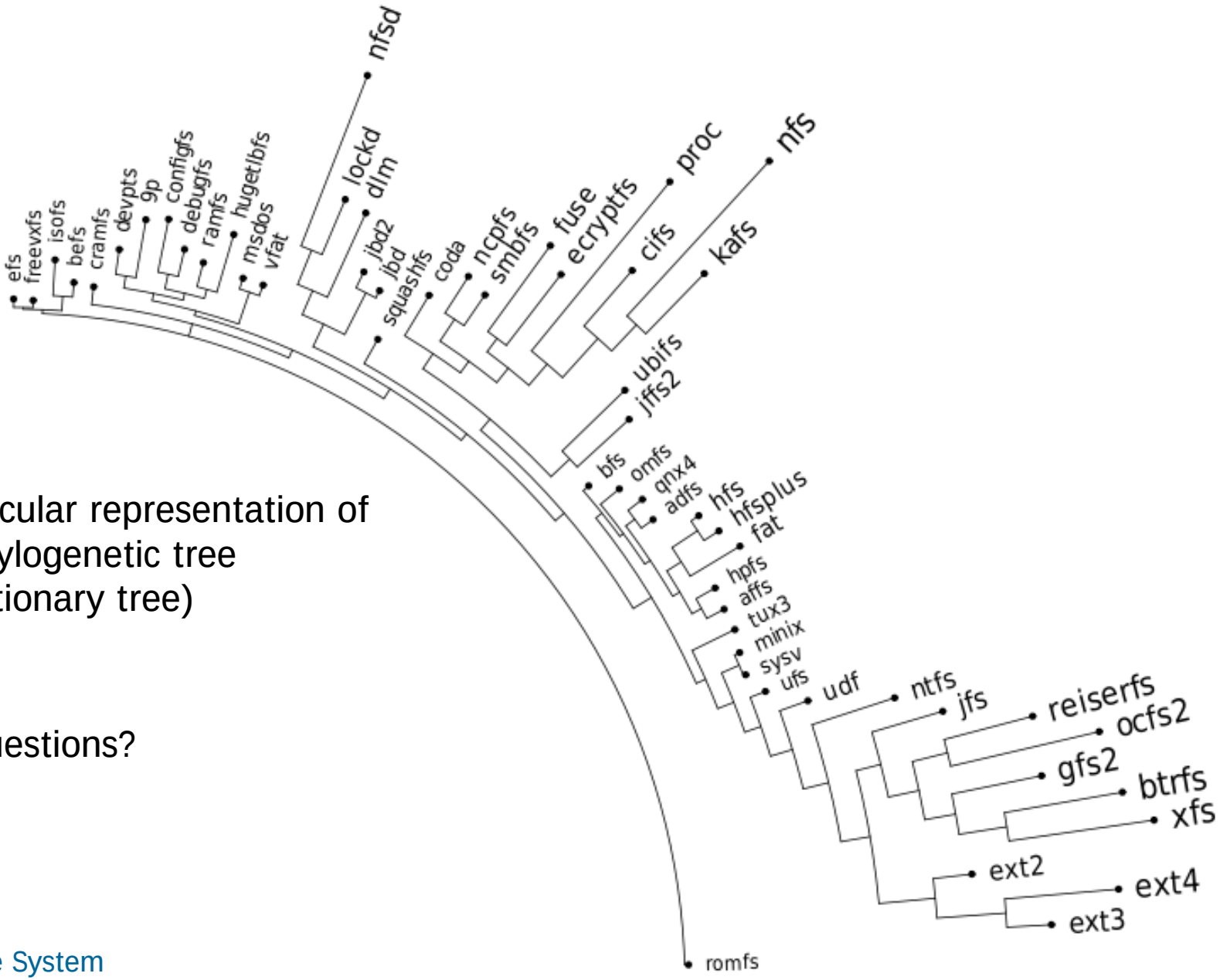
Visual Linux File Systems



Visual Linux File Systems



Visual Linux File Systems



90° circular representation of the phylogenetic tree (evolutionary tree)

Any questions?

Remember

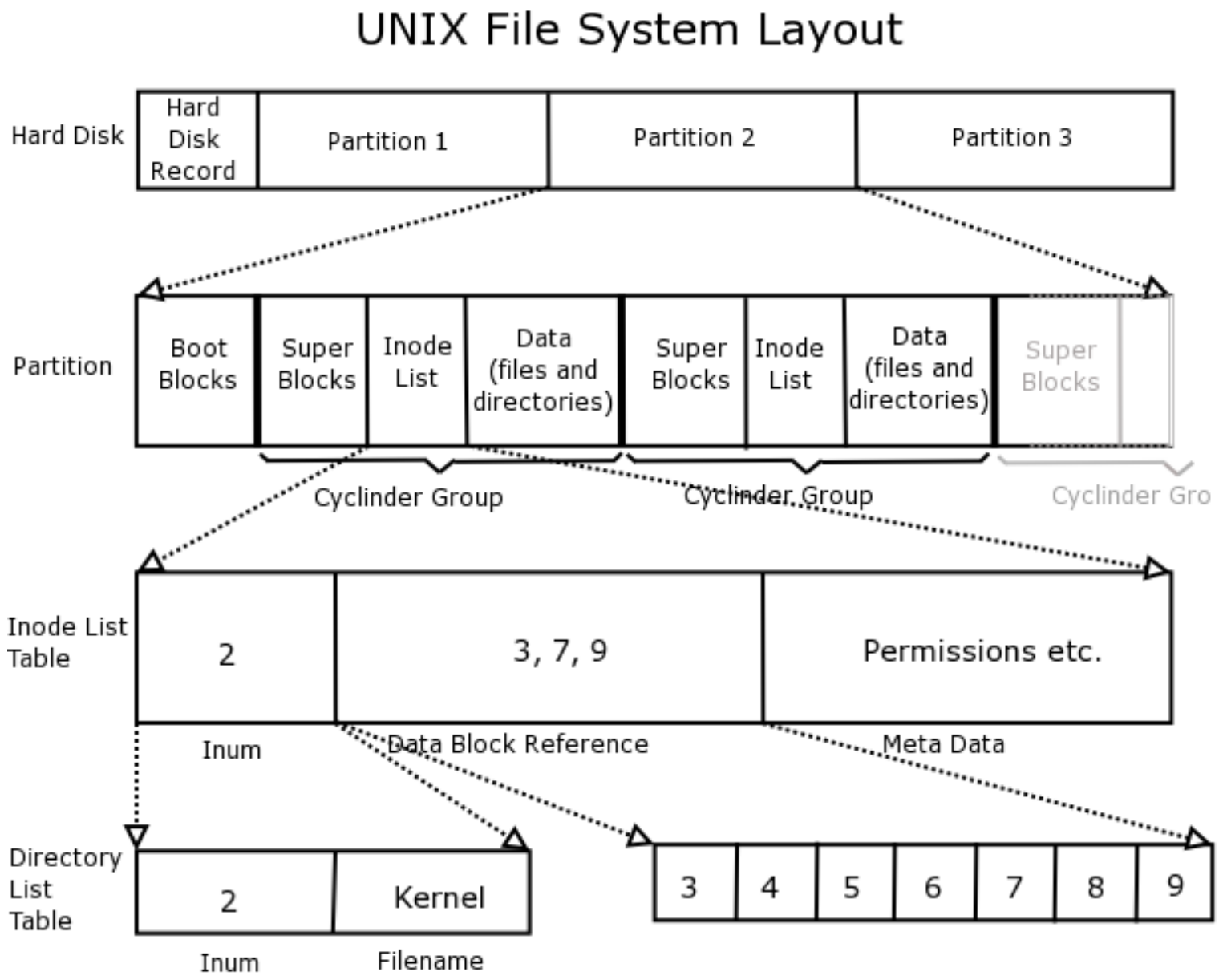
- File systems: FAT32 NTFS EXT4 SWAP
- Block vs Character devices
- Journaling tracks changes during last file sessions
- ACID (atomicity, consistency, isolation, durability)
- Joliet combined Romeo and ISO 9660
- Partitions can contain different file and operating systems
- RAID (Redundant Array of Inexpensive Disks)
  - RAID 0 Striping
  - RAID 1 Mirroring
  - RAID 5 Parity across disks
- JBOD – Just a bunch of disks
- NAS - Network Attached Storage vs SAN Storage Area
- Tera Peta Exa Zetta Yotta

Booting Time

Boot Loader

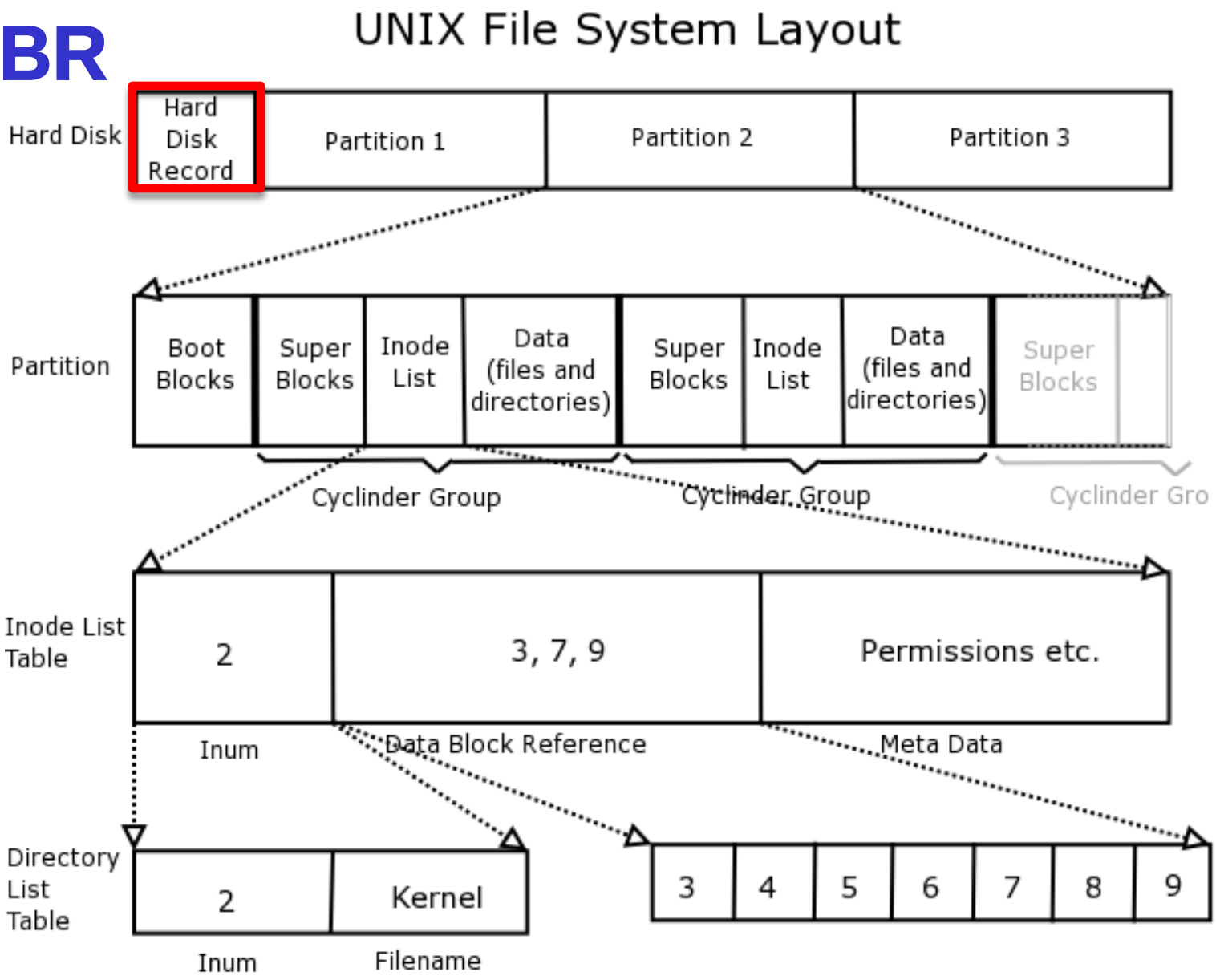
- At boot time the BIOS looks for instructions on loading the OS from drives first sector
- The master boot record (MBR) is on the first sector of the hard drive
- Depending on the boot loader, additional files may be stored/read from a partition on the drive
- The boot loader begins to start the operating system, and is not used again until the next boot
- Most current OS support a GUID Partition Table (GPT) as a standard layout for the partition table
- fdisk does not support GPT, 64-Windows does

UNIX File System Overview



UNIX File System Overview

MBR



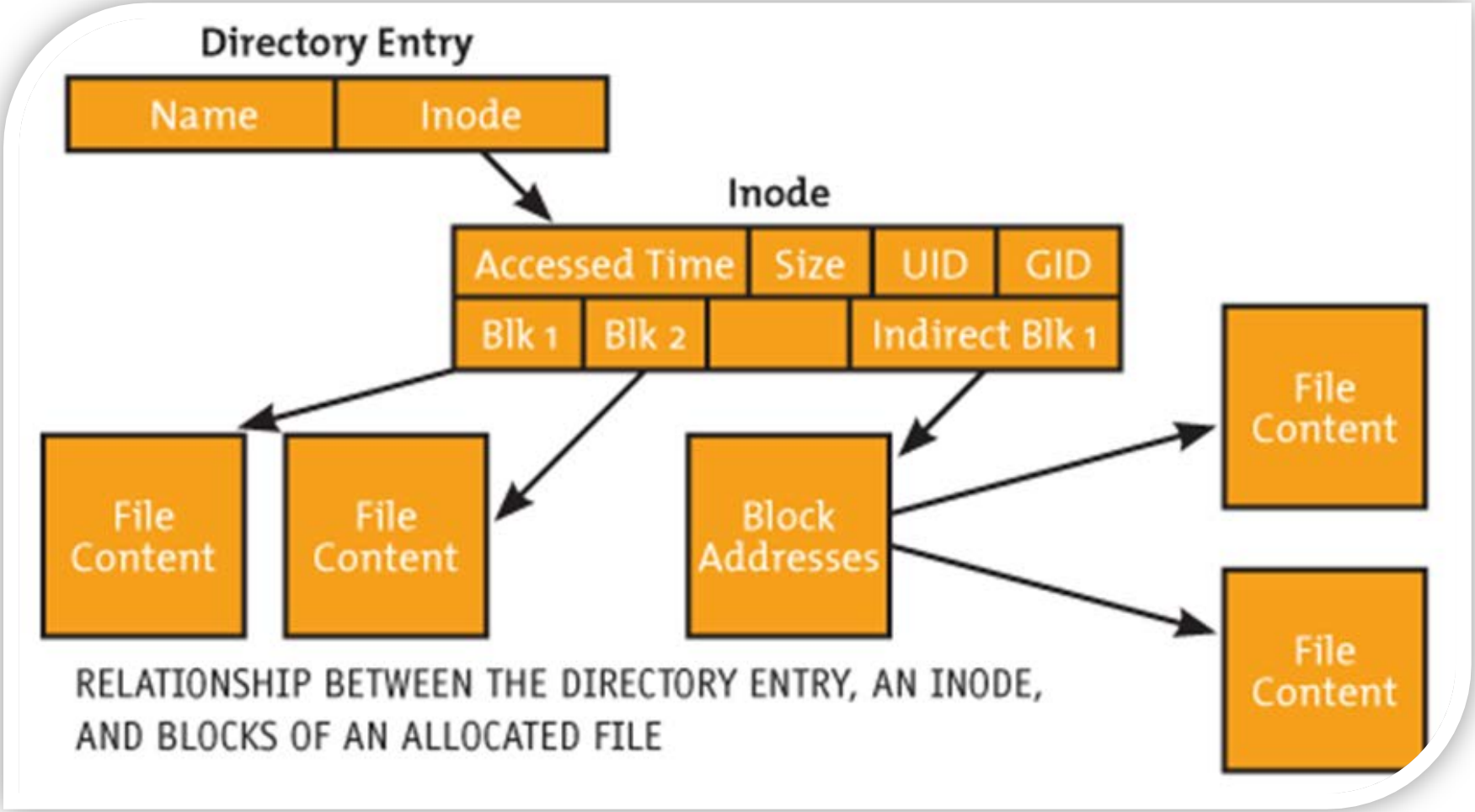
Linux File System

- The first 1024 bytes contains the boot code
- First data structure is the Superblock (~ Boot Record), located 1024 bytes from the beginning of the drive. Unlike the boot record, there is no boot code in the Superblock.
- Superblock contains:
  - Total number of blocks in whole file system
  - Number of blocks per block group
  - Number of reserved blocks before 1st block group
  - Total number of inodes in system
  - Number of inodes per block group
  - Number of sectors per block

Superblock

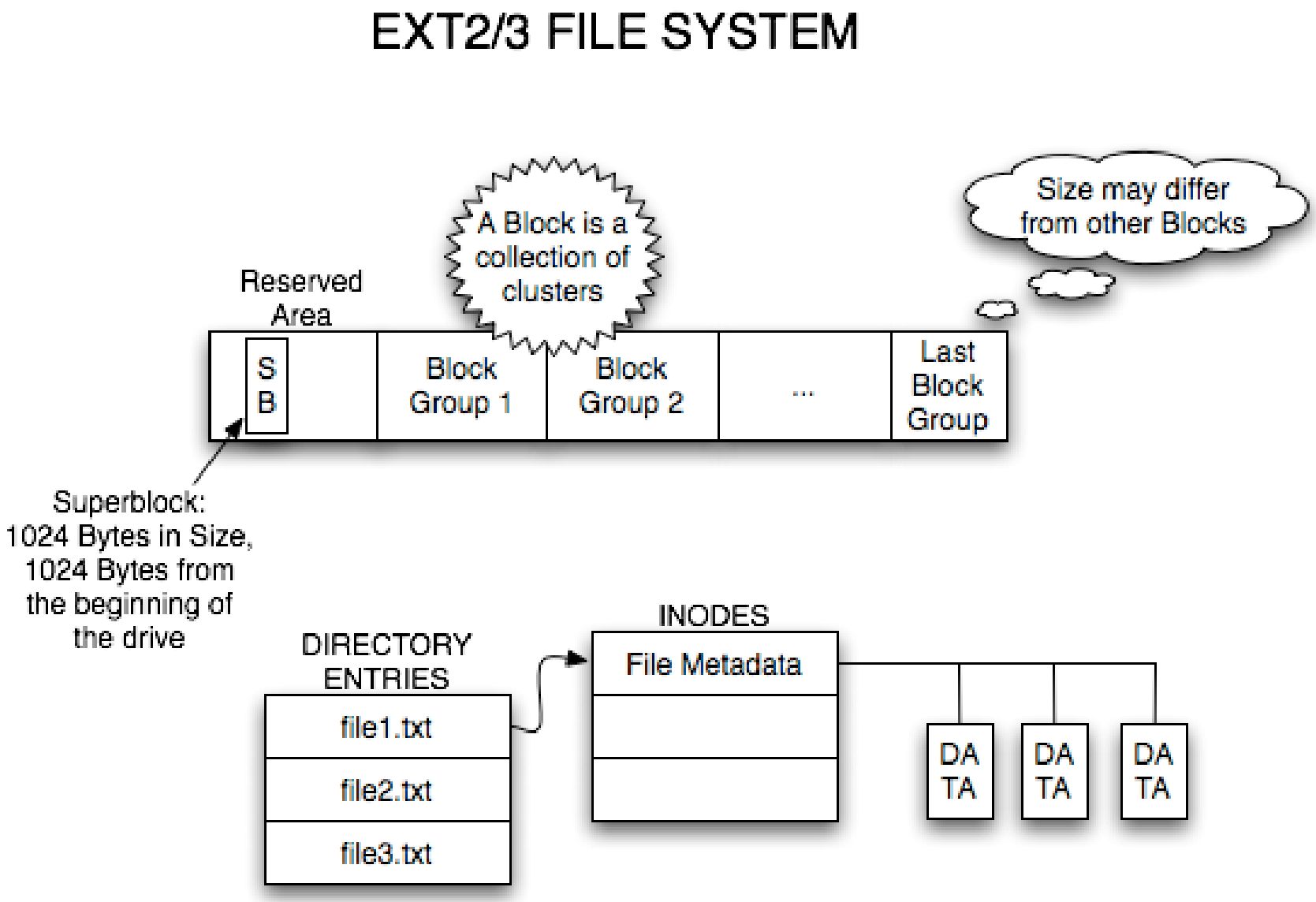
- A structure that represents a file system
- Includes information to manage the file system during operation
- Includes file system name (such as ext2)
- Size of the file system and its state
- A reference to the block device
- Metadata information (such as free lists)
- Typically stored on the storage medium
- Linux boot record called a Superblock. Linux divides the drive up into block groups, often a copy of Superblock at beginning of block groups

inodes



RELATIONSHIP BETWEEN THE DIRECTORY ENTRY, AN INODE, AND BLOCKS OF AN ALLOCATED FILE

Linux File System



Linux File Systems

ext4 (Fourth Extended Filesystem)

- **Default file system for Ubuntu**
- Supports volume sizes up to 1 exbibyte (EiB)
- Supports file sizes up to 16 tebibytes (TiB)
- Backward compatible with ext2 and ext3
- Extents replace block mapping
  - Extent: range of contiguous physical blocks
- Uses journal checksums
- Nanosecond timestamps
- Faster file system checking

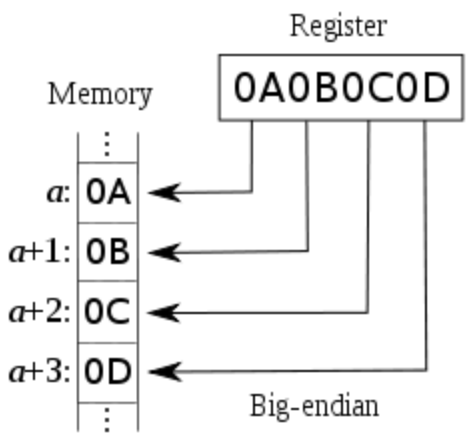
Linux File Systems

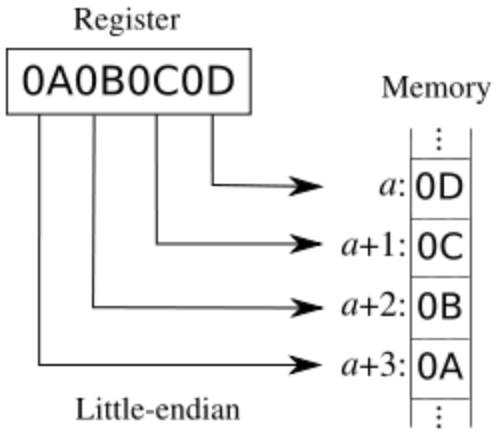
ext4

- Divides storage into array of logical blocks
- Reduces accounting overhead
- ↑ throughput by forcing larger transfer sizes
- Block allocator tries to keep each file's blocks in same group – less fragmentation
- Block size typically 4KB
- Same size as x86 pages – coincidence?
- Each group 32,786 blocks = 128MB
- Fields written to disk in little-endian order
- Fields in jbd2 journal written big-endian order

Endianness

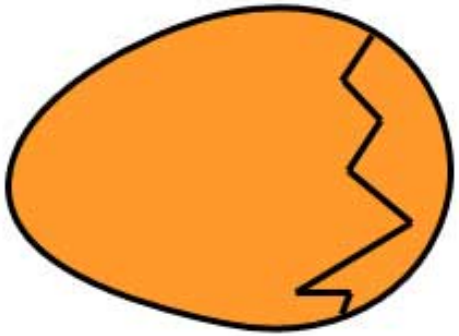
- A big-endian machine stores the most significant byte first, and a little-endian machine stores the least significant byte first.



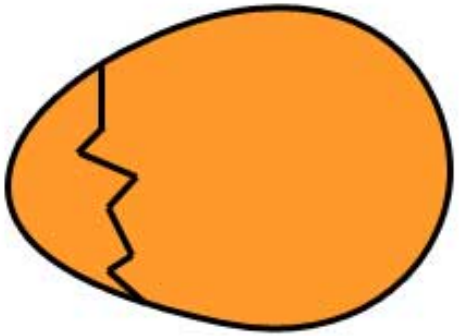


Endians

- Word origin: Jonathan Swift's *Gulliver's Travels*. In the Lilliput kingdom, a decree by the King that eggs be broken from the little side, caused a rebellion by a faction called the Big Endians, who were used to breaking their eggs from the bigger end.



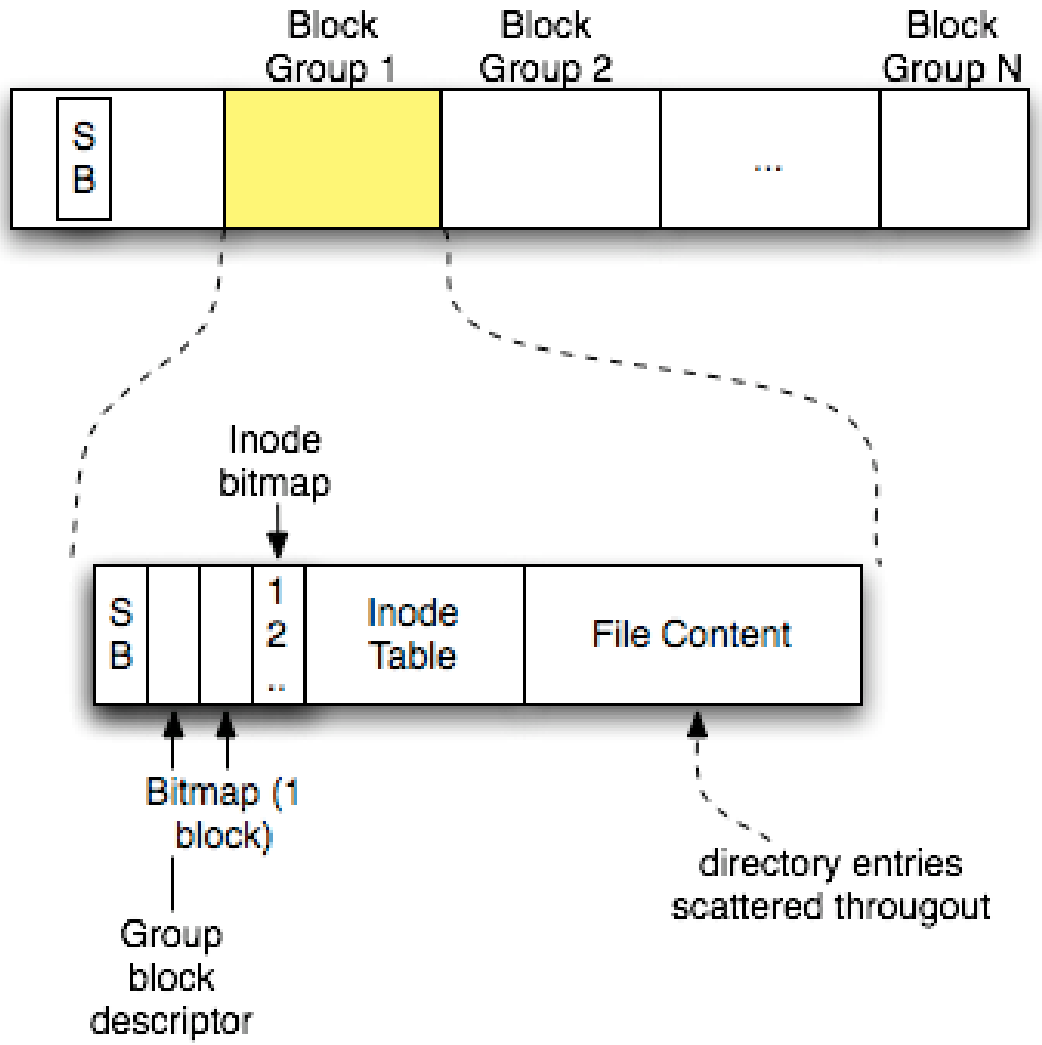
BIG ENDIAN - The way people always broke their eggs in the Lilliput land



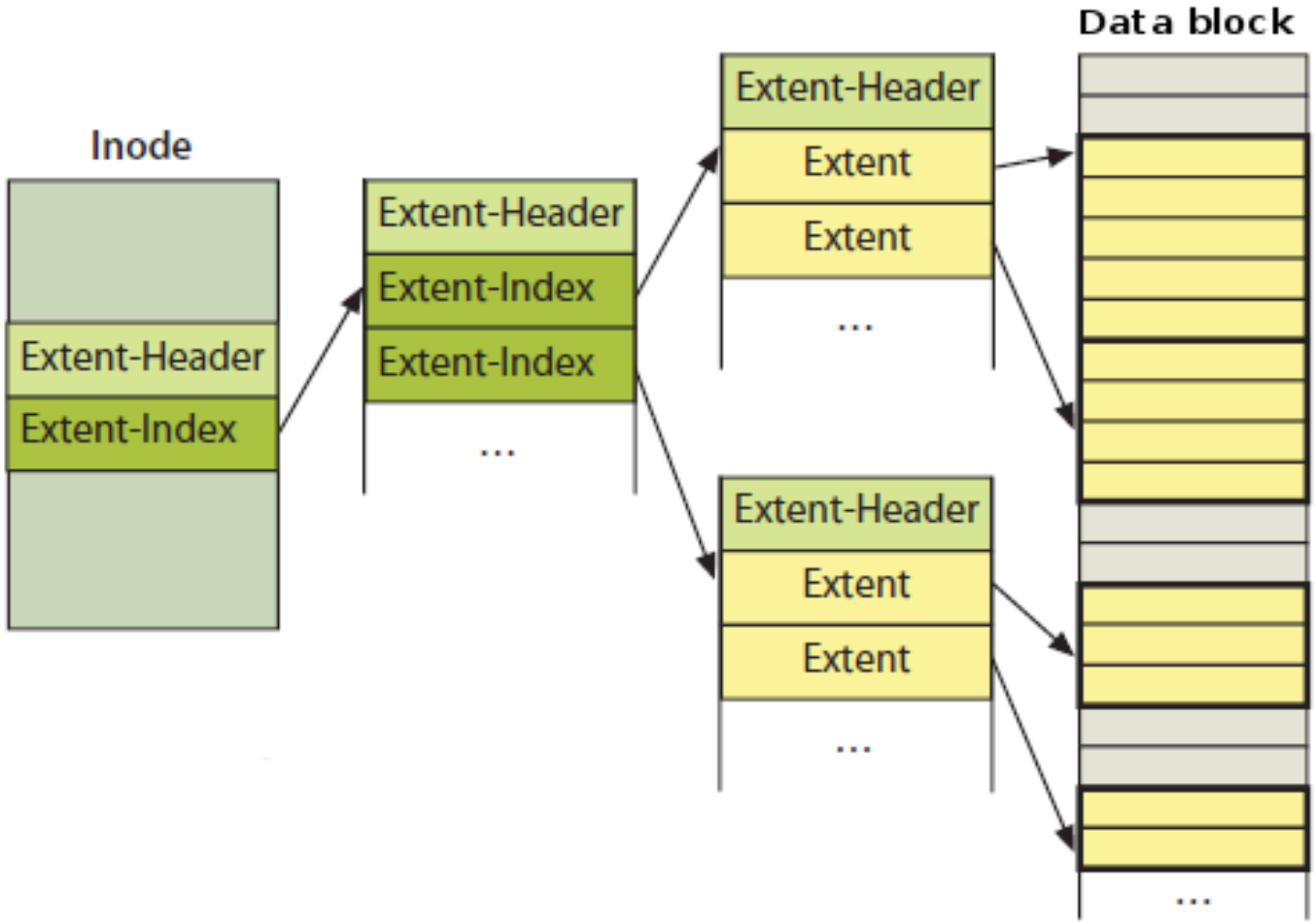
LITTLE ENDIAN - The way the king then ordered the people to break their eggs

Linux File System

WHAT'S IN A BLOCK GROUP?



ext4 extent tree



- ext4 reduces management overhead for large files
- ext4 can help prevent fragmentation

XFS

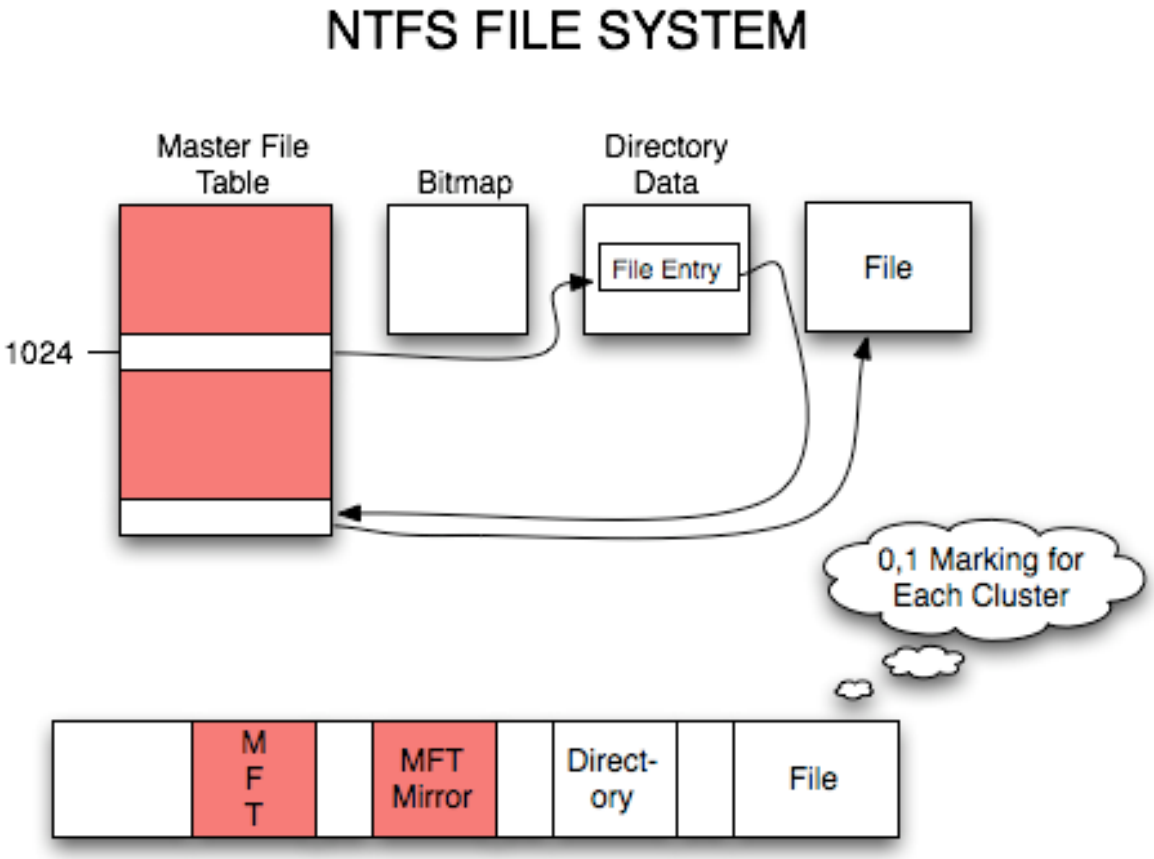
- High-performance journaling file system
- Created by Silicon Graphics for their IRIX OS
- Ported to 2.4 Linux kernel
- 64-bit – supports 8 exbibytes
- Proficient at handling large files
- Offers smooth data transfers for real-time applications, video-streaming
  - Direct I/O provision for high throughput
  - Guaranteed-rate I/O API

ZFS

- ZFS: file system and logical volume manager
- Designed focus on data integrity (no bit rot)
- Developed by Sun Microsystems (now Oracle)
- Stores up to 256 quadrillion zettabytes
- Continuous integrity checking & automatic repair
- Built on top of virtual storage pools
- Default OpenSolaris file system
- Supports deduplication
- Supports snapshots

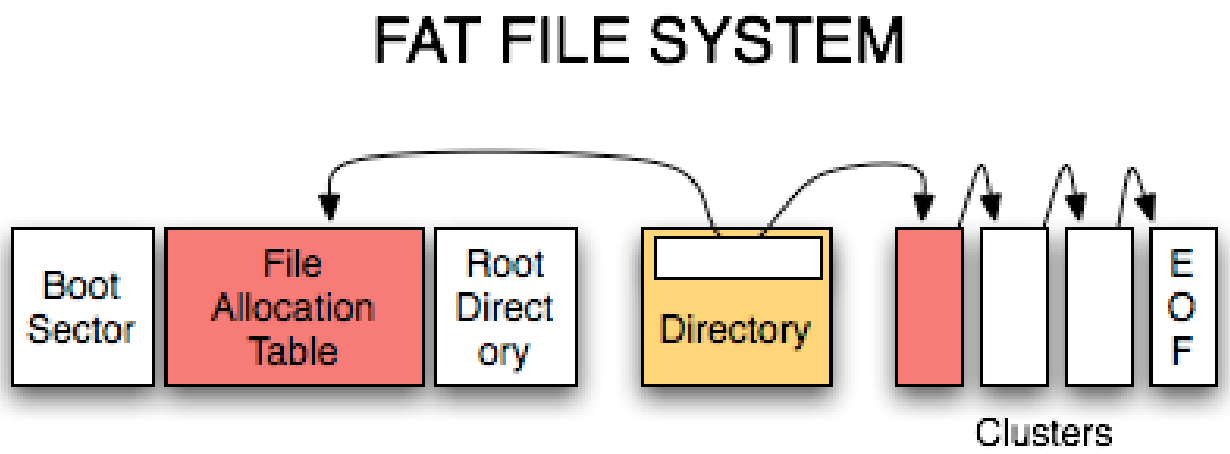
NTFS File System

- Master File Table at the front of the drive, with a mirror file in the middle of the drive. Only the first directory contains critical information.



Fat32 File System

- File System information (cluster size, sector size, etc.) is located in the boot sector. Root directory is in the data area, may grow as needed (and may be fragmented), starts in cluster 2.
- Each directory entry points to a starting cluster and to a place in the FAT where the cluster chain for the rest of the file is located.



Filesystem Classes

- Filesystems can be categorized into classes:
- **Image** - special filesystem that presents the modules in the image and is always present
- **Block** - traditional filesystems that operate on block devices like hard disks and CD-ROM drives.
- **Flash** - Nonblock-oriented filesystems designed explicitly for the characteristics of flash memory devices. NOR devices - FFS3 NAND - ETFS.
- **Network** - provide network file access to the filesystems on remote hosts. NFS and CIFS (SMB)
- **Virtual** - resource manager that sits in front of other file systems

Flash

- Flash memory - non-volatile storage chip
- Can be electrically erased and reprogrammed
- Developed from EEPROM (electrically erasable programmable read-only memory)
- High density NAND read/write in blocks or pages
  - memory cards
  - USB flash drives
  - solid-state drives
- NOR type allows a single byte read/write

Flash

- Flash memory allows a limited number of erase cycles on a flash block before the block will fail. This number can be as low as 100,000.
- If the majority of files in flash are static, this will cause the remaining blocks containing dynamic data to wear at a dramatically increased rate.
- Each read operation within a NAND flash block weakens the charge maintaining the data bits. Most devices support about 100,000 reads before there's danger of losing a bit.

BOOTING

Booting

Secondary boot loader allows intervention:

- Select alternate kernel to load
- Modify boot parameters
  - single user, extensions, ...
- USE PASSWORD SECURITY HERE!!!
- Depending on security policy-
  - Secure the boot process so users cannot boot alternate media, an alternate kernel, etc.

Secondary Boot Loaders

- Secondary boot loaders have become smarter
- Can locate/load/start several OS kernels
- OS vendors published information about
  - How their boot loaders work
  - Where they look for information required to load the kernel, boot the system

- A good example of this type of boot loader is the GRand Unified Boot loader (GRUB).

GRUB

```
# /etc/grub.conf generated by anaconda    (Installer)
timeout=10
splashimage=(hd0,1)/grub/splash.xpm.gz
password --md5 $1$ÕpîÁÜdþï$J08sMAcfyWW.C3soZpHkh.
title Red Hat Linux (2.6.37-3custom)
        root (hd0,1)
        kernel /vmlinuz-2.6.37-3custom ro root=/dev/hda5
        initrd /initrd-2.6.37-3.img
title Red Hat Linux (2.4.18-3) Emergency kernel (no afs)
        root (hd0,1)
        kernel /vmlinuz-2.4.18-3 ro root=/dev/hda5
        initrd /initrd-2.4.18-3.img
title Windows 7 Professional
        rootnoverify (hd0,0)
        chainloader +1
```

Linux Loader (LILO)

```
# sample /etc/lilo.conf
boot = /dev/hda
delay = 40
password=SOME_PASSWORD_HERE
default=vmlinuz-stable
vga = normal
root = /dev/hda1
image = vmlinuz-2.5.99
   label = net test kernel
   restricted
image = vmlinuz-stable
   label = stable kernel
   restricted
other = /dev/hda3
   label = Windows 2000 Professional
   restricted
   table = /dev/hda
```

OS Kernel

- The kernel loads, starts, and probes devices
  - System busses probed for devices to build the device tree.
- Once the device tree is available, the kernel parses it, and each device is probed to determine if it is operational, (and if so, the driver module is loaded into the kernel)
  - This search for memory and devices is sometimes referred to as auto-configuration

Good Post-Boot Commands

- dmesg (display message) command
  - Displays the message buffer of the kernel
- lspci
  - Displays info on all PCI bus devices
- lsusb
  - Displays info on all USB devices
- uname –a
  - Displays computer name, kernel, CPU
  - Linux hatter 2.6.37.6-smp #2 SMP
  - Sat Apr 9 23:39:07 CDT 2011
  - i686 Intel(R) Pentium(R) 4 CPU 2.00GHz GenuineIntel
  - GNU/Linux

dmesg

- Post boot – dmesg is all about the boot
- Afterwards – further kernel messages
- *try sudo dmesg | more*

```
[2392031.089261] net_ratelimit: 3 callbacks suppressed
[2392031.089269] TCP: Possible SYN flooding on port 110. Sending cookies.
[2392061.125806] TCP: Possible SYN flooding on port 110. Sending cookies.
[2392061.133469] TCP: Possible SYN flooding on port 110. Sending cookies.
[2392061.143435] TCP: Possible SYN flooding on port 110. Sending cookies.
[2392061.156062] TCP: Possible SYN flooding on port 110. Sending cookies.
[4038057.977051] TCP: Possible SYN flooding on port 25. Sending cookies.
[4038060.876026] TCP: Possible SYN flooding on port 25. Sending cookies.
```

dmesg excerpts

Enabling APIC mode: Flat. Using 1 I/O APICs
CPU0: Intel Pentium 4 CPU 2.00GHz stepping 04
Total of 1 processors activated (3983.75 BogoMIPS)
ACPI: PCI Interrupt Link IRQs 3 4 *5 6 7 9 10 11 14 15
PCI: Using ACPI for IRQ routing
usbcore: registered new interface driver usbfs
usbcore: registered new interface driver hub
usbcore: registered new device driver usb
serial8250: ttyS0 at I/O 0x3f8 (irq = 4) is a 16550A
ACPI: Thermal Zone [THRM] (30 C)
ACPI: Fan [FAN] (on)

OS Kernel

UNIX Run Levels

- All flavors of UNIX/Linux use similar foundations for the system run modes
- There are basically two run modes:
  - Single user (aka maintenance mode), and
  - Multi-user
    - There may be several forms of the multi-user mode (with services, without services, etc.) (with GUI or with CLI) on any given UNIX OS
    - Debian/Ubuntu desktop dropped CLI option
    - MacOS X kernel has same set of run levels

Linux Runlevel Example

Table 1: Linux Runlevels	
0	Halt
1	Single-user mode
2	Multiuser, without NFS
3	Full multiuser mode
4	Unused
5	X11 (GUI)

OS Kernel

Windows Run Levels

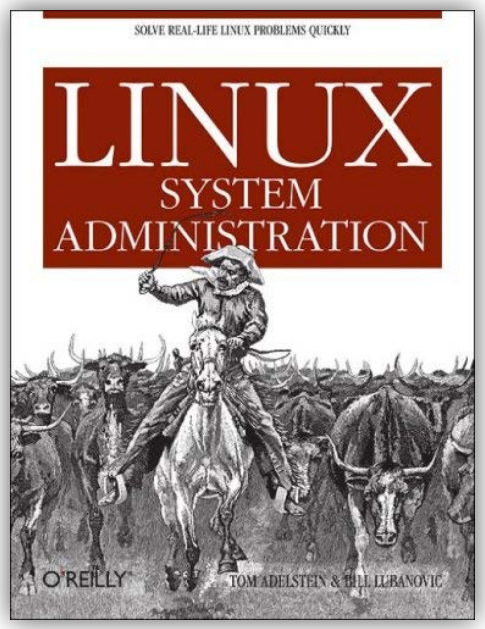
- Windows has a limited set of run levels
  - Multi-user
  - Safe Mode
  - Safe mode with networking
    - Typically used to repair damaged system

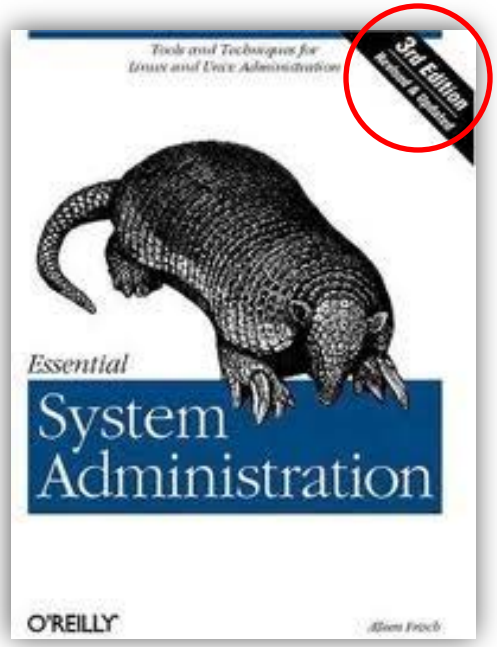
Caveats

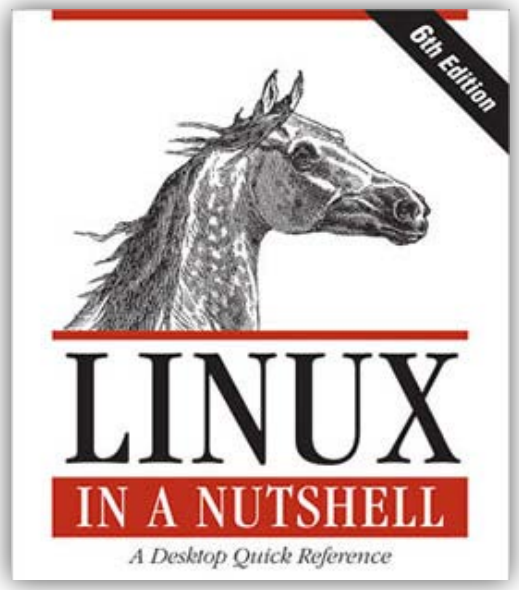
- e4defrag to defragment files
- Run 64-bit Linux guest OS on Intel 64?
  - Was called EM64T
  - e.g. Core 2, Core i3/5/7, Xeon, Atom
  - Any non-lamed CPU since 7/2006
- Enable "Execute Disable" in the host BIOS
- Helps ensure Linux guest OS runs w/o interruption

Additional References

- http://tldp.org/LDP/sag/html/index.html
- http://www.yolinux.com/TUTORIALS/LinuxTutorialSysAdmin.html







Remember

- File systems: FAT32 NTFS EXT4 SWAP
- Block vs Character devices
- Journaling tracks changes during last file sessions
- ACID (atomicity, consistency, isolation, durability)
- Joliet combined Romeo and ISO 9660
- Partitions can contain different file and operating systems
- RAID (Redundant Array of Inexpensive Disks)
  - RAID 0 Striping
  - RAID 1 Mirroring
  - RAID 5 Parity across disks
- JBOD – Just a bunch of disks
- NAS - Network Attached Storage vs SAN Storage Area
- Tera Peta Exa Zetta Yotta

Remember

- MBR: Master Boot Record
- Superblock is a boot record
- Big-endian machine stores most significant byte first
- Little-endian machine stores least significant byte first
- Run level Linux (single, assorted multi-user levels)
- Run level Windows (normal, safe)