CS615 - Aspects of System Administration Filesystems, Disks, Storage

Department of Computer Science Stevens Institute of Technology Jan Schaumann

jschauma@stevens.edu

http://www.cs.stevens.edu/~jschauma/615A/

http://www.cs.stevens.edu/~jschauma/615/s14-hw1.html

http://www.cs.stevens.edu/~jschauma/615/s14-hw1.html

Running an instance:

http://www.cs.stevens.edu/~jschauma/615/s14-hw1.html

Save yourself some typing:

\$ instance <AMI-ID>

http://www.cs.stevens.edu/~jschauma/615/s14-hw1.html

Make it permanent:

http://www.cs.stevens.edu/~jschauma/615/s14-hw1.html

ssh to an instance:

\$ ssh -i ~/.ec2/stevens.pem root@<mumble>.compute-1.amazonaws.com

http://www.cs.stevens.edu/~jschauma/615/s14-hw1.html

Let's save ourselves some typing:

Filesystems, Disk, Storage

```
$ instance ami-35eb835c
[...]
$ aws ec2 describe-instances
[...]
$ ssh root@ec2-54-198-235-45.compute-1.amazonaws.com
```

```
# uname -a
SunOS domU-12-31-39-0F-1C-BB.compute-1.internal 5.11 omnios-33fdde4 i86pc
i386 i86xpv Solaris
#
```

```
# netstat -na | more
[...]
TCP: IPv4
  Local Address Remote Address Swind Send-Q Rwind Recv-Q
State
127.0.0.1.4999
                                    0
                                         0 128000
                                                     O LISTEN
                     * . *
    *.111
                                         0 128000
                                                     O LISTEN
                     * . *
                                          0 128000
                                                     0 IDLE
    *.*
                     *.*
                                          0 128000
                                                     O LISTEN
    *.111
                     *.*
                                          0 128000
                                                     0 IDLE
    *.*
                     * . *
                                         0 128000
                                                     O LISTEN
    *.46457
                     *.*
    *.55986
                                         0 128000
                                                     O LISTEN
                     *.*
    *.22
                                         0 128000
                                                     O LISTEN
                     *.*
O ESTABLISHED
[\ldots]
```

```
# man df
[...]
# df
[...]
# df -hT
[...]
# df -i
[...]
# df -a
[...]
# mount
[...]
```

```
# format
[...]
format> verify
[...]
# zpool list
[...]
# zfs list
[...]
```

Filesystems, Disks, Storage

- basic disk concepts
- basic filesystem concepts
- file systems

Topics covered

- basic disk concepts
 - storage models
 - disk interfaces
 - physical disk structure
 - partitions
- basic filesystem concepts
- file systems

Topics covered

- basic disk concepts
 - storage models
 - disk interfaces
 - physical disk structure
 - partitions
- basic filesystem concepts
 - RAID
 - logical volume managment
 - device formatting
- file systems

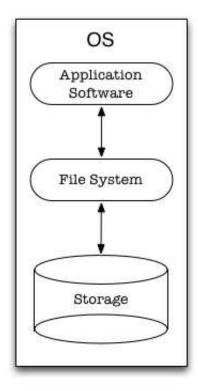
Topics covered

- basic disk concepts
 - storage models
 - disk interfaces
 - physical disk structure
 - partitions
- basic filesystem concepts
 - RAID
 - logical volume managment
 - device formatting
- file systems
 - the UNIX filesystem or Berkeley Fast File System (FFS)

Basic Disk Concepts

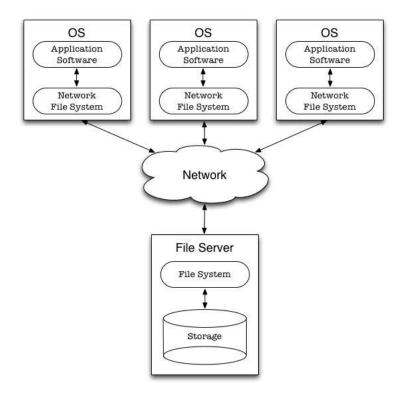
Storage Models

Direct Attached Storage (DAS)



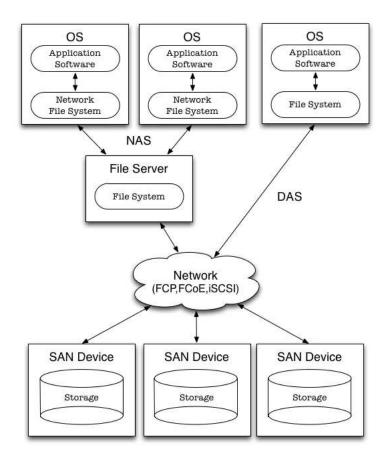
ssh lab 'df -hT /'

Network Attached Storage (NAS)

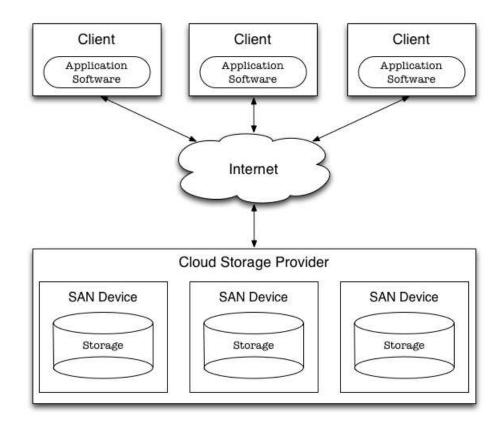


ssh lab 'df -hT /home/\$(whoami)'

Storage Area Networks (SAN)



Cloud Storage (Examples: EBS, S3)



Basic Disk Concepts: Storage Models: Cloud Storage

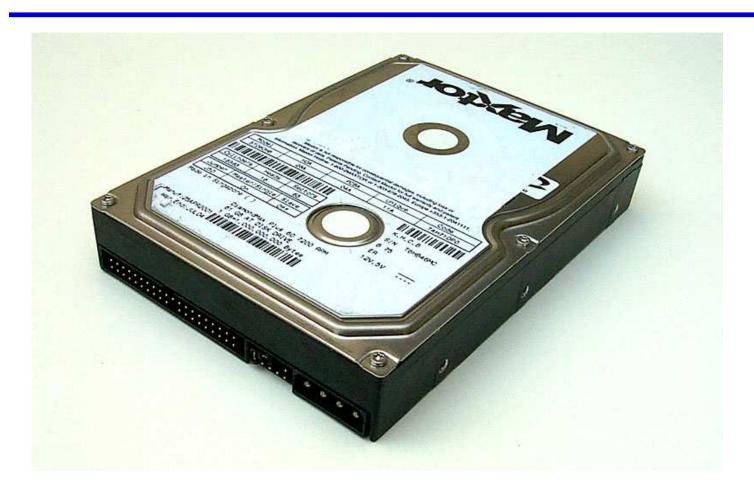
```
$ aws ec2 describe-instances
[...]
/dev/sda1 ebs None paravirtual
BLOCKDEVICEMAPPINGS /dev/sda
EBS 2014-01-25T20:18:19.000Z True attached vol-a0d000d6
[...]
```

Basic Disk Concepts: Storage Models: Cloud Storage

```
$ aws ec2 create-volume --size 1 --availability-zone us-east-1d
[...]
$ aws ec2 attach-volume --volume-id vol-9d3aeaeb --instance-id \
        i-dd74f0f3 --device /dev/sdh
[...]
$ ssh <hostname>
# format
format> fdisk
format> verify
format> label
# newfs /dev/rdsk/c1t2160d0s0
[...]
# mount /dev/dsk/c1t2160d0s0 /mnt
# df -Th /mnt
[...]
# fstyp -v /dev/rdsk/c1t2160d0s0 | more
[...]
```

Basic Disk Concepts

Disk Devices















Basic Disk Concepts

Disk Interfaces

Basic Disk Concepts: Disk Interfaces: SCSI



Basic Disk Concepts: Disk Interfaces: ATA



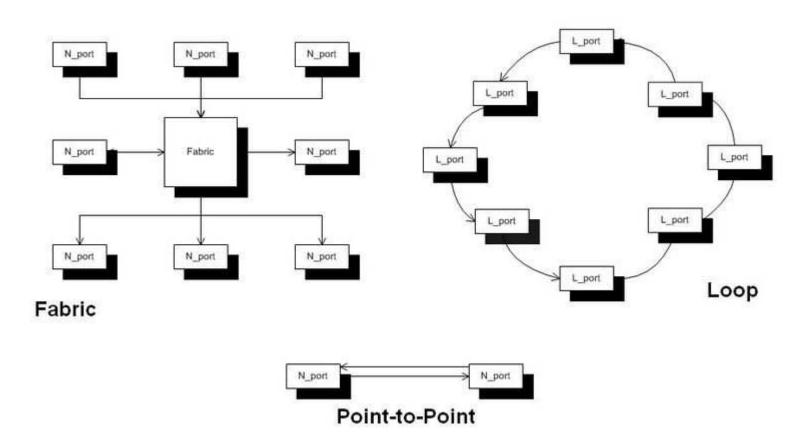
Basic Disk Concepts: Disk Interfaces: ATA



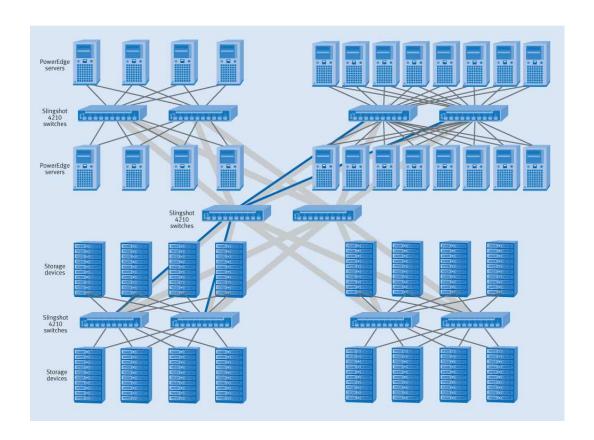
Basic Disk Concepts: Disk Interfaces: Fibre Channel











Basic Disk Concepts: Disk Interfaces: SANs

- ATA over Ethernet (AoE):
 - create low-cost SAN
 - ATA encapsulated into Ethernet frames
- Fibre Channel over Ethernet (*FCoE*):
 - consolidate IP and FC/SAN networks
 - FC encapsulated into Ethernet frames
- *oE:
 - no TCP/IP overhead
 - restricted to a single Layer 2 network
 - no inherent security features
- iSCSI
 - SCSI encapsulated in TCP/IP packets

Basic Disk Concepts

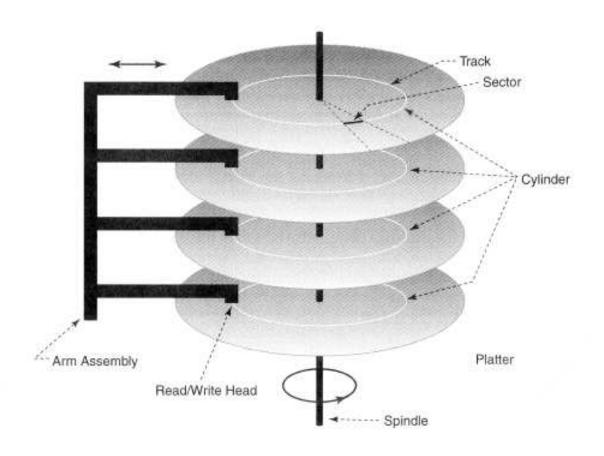
Physical Disk Structure







Basic Disk Concepts: Physical Disk Structure



Basic Disk Concepts: Physical Disk Structure

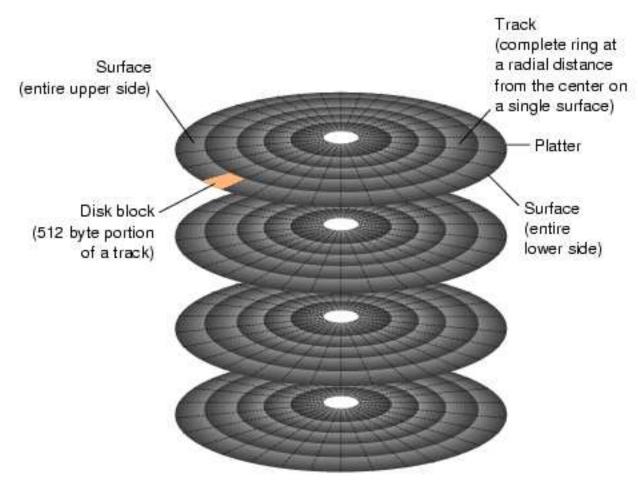
Hard drive performance determined by:

- seek time
- rotational latency
- internal data rate
- a few other negligible factors (external data rate, command overhead, access time, etc.)

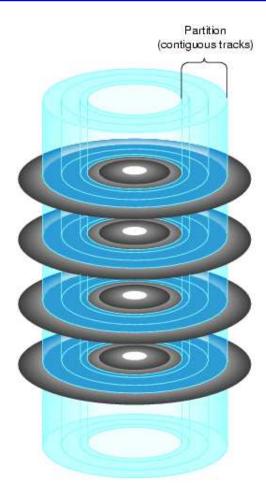


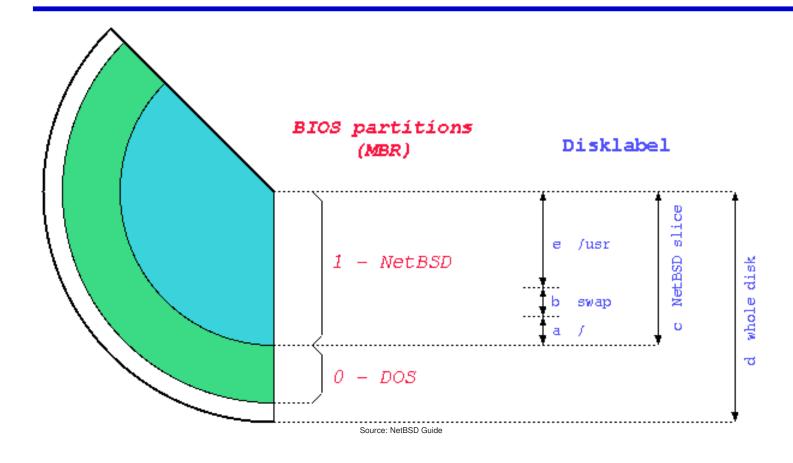
Basic Disk Concepts

Partitions



Source: SGI Techpubs





NetBSD example (from disklabel(8))

Partition 'a': /

Partition 'b': swap Partition 'e': /home

```
#
        size
               offset
                       fstype [fsize bsize cpg/sgs]
   20972385
                  63
                      4.2BSD
                               4096 32768
                                         1180
                                               # (Cyl. 0*- 20805)
a:
    1048320 20972448
                                               # (Cyl. 20806 - 21845)
b:
                        swap
                                               # (Cyl. 0*- 77519)
  78140097
                  63
                      unused
                                 0
                                       0
c:
                                               # (Cyl. 0 - 77519)
  78140160
                                 0
d:
                   0
                      unused
                                       0
   56119392 22020768
                     4.2BSD
                               4096 32768 58528
                                               # (Cyl. 21846 - 77519)
e:
```

NetBSD example (from disklabel(8))

Partition 'a': / 10 GB

Partition 'b': swap

Partition 'e': /home 26 GB

```
#
        size
               offset
                       fstype [fsize bsize cpg/sgs]
   20972385
                  63
                      4.2BSD
                               4096 32768
                                          1180
a:
                                               # (Cyl. 0*-20805)
   1048320 20972448
                                               # (Cyl. 20806 - 21845)
b:
                        swap
                                               # (Cyl. 0*- 77519)
  78140097
                  63
                      unused
                                 0
                                       0
c:
                                               # (Cyl.
  78140160
                                 0
                                                           0 - 77519)
d:
                   0
                      unused
                                       0
   56119392 22020768
                     4.2BSD
                               4096 32768 58528
                                               # (Cyl. 21846 - 77519)
e:
```

Solaris example (from format(1m)):

Current partition table (original):

Total disk cylinders available: 38758 + 2 (reserved cylinders)

Part	t Tag	Flag	Cyli	nders	Size	Block	S
0	root	wm	3 -	3764	3.62GB	(3762/0/0)	7584192
1	swap	wu	3765 -	4364	590.62MB	(600/0/0)	1209600
2	backup	wm	0 -	38757	37.26GB	(38758/0/0)	78136128
3	unassigned	wm	0		0	(0/0/0)	0
4	unassigned	wm	0		0	(0/0/0)	0
5	unassigned	wm	0		0	(0/0/0)	0
6	unassigned	wm	0		0	(0/0/0)	0
7	home	wm	4365 -	38757	33.06GB	(34393/0/0)	69336288
8	boot	wu	0 -	. 0	0.98MB	(1/0/0)	2016
9	alternates	wu	1 -	. 2	1.97MB	(2/0/0)	4032

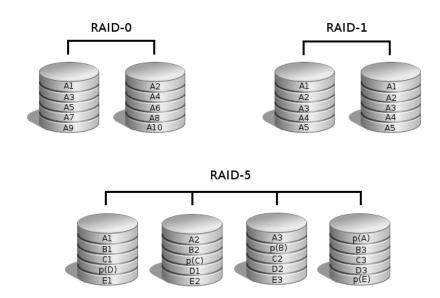
Linux example (from fdisk(8)):

Disk /dev/sda: 80.0 GB, 80000000000 bytes 255 heads, 63 sectors/track, 9726 cylinders Units = cylinders of 16065 * 512 = 8225280 bytes

Device	Boot	Start	End	Blocks	Id	System
/dev/sda1	*	1	33	265041	83	Linux
/dev/sda2		34	9726	77859022+	83	Linux

Basic Disk and Filesystem Concepts: RAID and Logical Volumes

- allow file systems to be larger than the physical size of a disk
- inrease I/O performance when striped
- fault tolerant when mirrored or plexed

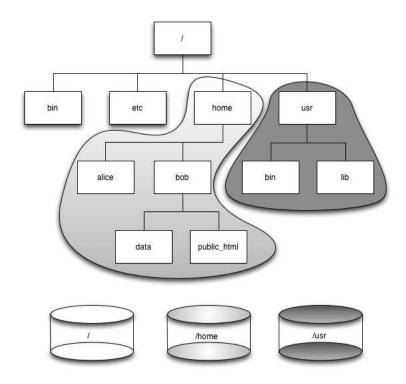


Hooray!

5 Minute Break

Filesystem Layout

All partitions – with the exception of the *root* (or /) partition – can be *mounted* anywhere in the filesystem hierarchy.



All partitions – with the exception of the *root* (or /) partition – can be *mounted* anywhere in the filesystem hierarchy.

The file /etc/fstab (see fstab(5)) specifies which disks / partitions to mount where:

```
/dev/wd0a
                  ffs rw 1 1
/dev/cgd1a
                  swap sw 0 0
          none
          /home ffs rw 1 2
/dev/cgd0a
/ignoreme
          /tmp mfs rw,-b4096,-f512,-s262144 0 0
          /kern kernfs rw
kernfs
          /proc
                procfs rw, noauto
procfs
ptyfs
          /dev/pts ptyfs rw 0 0
```

All partitions – with the exception of the *root* (or /) partition – can be *mounted* anywhere in the filesystem hierarchy.

The file /etc/fstab (see fstab(5)) specifies which disks / partitions to mount where:

```
# /etc/fstab: static file system information.
# Use 'vol_id --uuid' 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>
                                                                <pass>
                                                        <dump>
                                                                0
                /proc
                                        defaults
proc
                                proc
# / was on /dev/sda2 during installation
LABEL=ROOT
                        ext3
                                errors=remount-ro,acl
# /boot was on /dev/sda1 during installation
                                                                2
LABEL=BOOT
                /boot ext3
                                defaults.acl
# swap was on /dev/sda5 during installation
UUID=9329ae83-289d-4c3d-8756-f707c4bbb312 none
                                                          swap
0
        0
/dev/scd0
                /media/cdrom0
                               udf, iso9660 user, noauto, exec, utf8 0
deathstar.phy.stevens-tech.edu:/export/nfs-sw/opt
                                                                     nfs ro, rsize=32768, intr, nolock 0 0
deathstar.phy.stevens-tech.edu:/export/srcit-dist
                                                     /mnt/srcit-dist nfs ro,rsize=32768,intr,nolock 0 0
corsario.cs.stevens-tech.edu:/export/people
                                                     /mnt/legacy/people nfs rw,rsize=32768,wsize=32768,intr,lock 0 0
corsario.cs.stevens-tech.edu:/export/faculty
                                                     /mnt/legacy/faculty nfs rw,rsize=32768,wsize=32768,intr,lock 0 0
```

All partitions – with the exception of the *root* (or /) partition – can be *mounted* anywhere in the filesystem hierarchy.

To see what filesystems are currently mounted, run mount (8):

```
/dev/wd0a on / type ffs (local)
/dev/cgd0a on /home type ffs (local)
mfs:276 on /tmp type mfs (synchronous, local)
kernfs on /kern type kernfs (local)
ptyfs on /dev/pts type ptyfs (local)
```

```
$ mount
/dev/sda2 on / type ext3 (rw,errors=remount-ro,acl)
tmpfs on /lib/init/rw type tmpfs (rw,nosuid,mode=0755)
proc on /proc type proc (rw,noexec,nosuid,nodev)
sysfs on /sys type sysfs (rw,noexec,nosuid,nodev)
varrun on /var/run type tmpfs (rw,nosuid,mode=0755)
varlock on /var/lock type tmpfs (rw,noexec,nosuid,nodev,mode=1777)
udev on /dev type tmpfs (rw,mode=0755)
tmpfs on /dev/shm type tmpfs (rw,nosuid,nodev)
devpts on /dev/pts type devpts (rw,noexec,nosuid,gid=5,mode=620)
fusectl on /sys/fs/fuse/connections type fusectl (rw)
1rm on /lib/modules/2.6.28-17-generic/volatile type tmpfs (rw,mode=755)
/dev/sda1 on /boot type ext3 (rw,acl)
securityfs on /sys/kernel/security type securityfs (rw)
automount(pid2623) on /home type autofs (rw,fd=4,pgrp=2623,minproto=2,maxproto=4)
deathstar.phy.stevens-tech.edu:/export/nfs-sw/opt on /opt type nfs (ro,rsize=32768,intr,nolock,addr=155.246.89.4)
deathstar.phy.stevens-tech.edu:/export/srcit-dist on /mnt/srcit-dist type nfs (ro,rsize=32768,intr,nolock,addr=155.246.89.4)
corsario.cs.stevens-tech.edu:/export/people on /mnt/legacy/people type nfs (rw,rsize=32768,wsize=32768,intr,lock,addr=155.246.89.20)
corsario.cs.stevens-tech.edu:/export/faculty on /mnt/legacy/faculty type nfs (rw,rsize=32768,wsize=32768,intr,lock,addr=155.246.89.20)
binfmt_misc on /proc/sys/fs/binfmt_misc type binfmt_misc (rw,noexec,nosuid,nodev)
deathstar.phy.stevens-tech.edu:/export/home/kamberov on /home/kamberov type nfs (rw,sync,intr,vers=3,sloppy,addr=155.246.89.4)
deathstar.phy.stevens-tech.edu:/export/home/mweiss on /home/mweiss type nfs (rw,sync,intr,vers=3,sloppy,addr=155.246.89.4)
deathstar.phy.stevens-tech.edu:/export/home/jschauma on /home/jschauma type nfs (rw,sync,intr,vers=3,sloppy,addr=155.246.89.4)
```

Some of the different kinds of filesystems:

Some of the different kinds of filesystems:

- "Regular" File Systems
- Journaling File Systems
- Network File Systems
- Various

The UNIX Filesystem

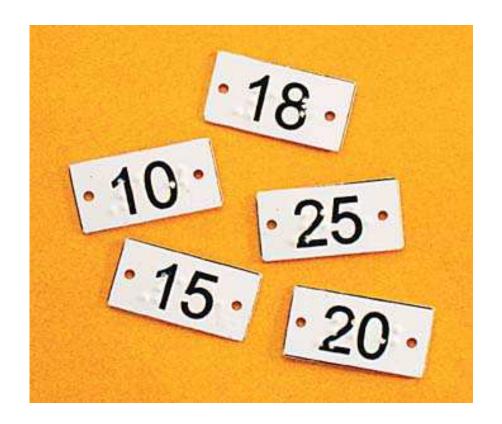
The filesystem is responsible for storing the data on the disk. So to read/write data, it needs to know in which physical blocks the actual data is located; ie how to map files to the disk blocks.

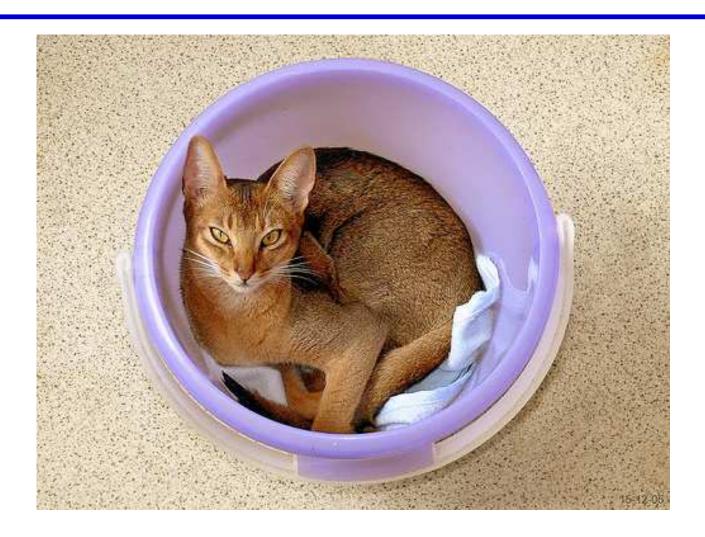
























The filesystem is responsible for storing the data on the disk. So to read/write data, it needs to know in which physical blocks the actual data is located; ie how to map files to the disk blocks.

Components of the Berkeley Fast Filesystem:

set of inode storage cells

df -i



The filesystem is responsible for storing the data on the disk. So to read/write data, it needs to know in which physical blocks the actual data is located; ie how to map files to the disk blocks.

Components of the Berkeley Fast Filesystem:

- set of inode storage cells
- set of scattered "superblocks"

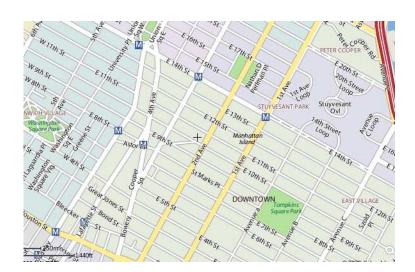
newfs -N /dev/rdsk/c1t2160d0s0



The filesystem is responsible for storing the data on the disk. So to read/write data, it needs to know in which physical blocks the actual data is located; ie how to map files to the disk blocks.

Components of the Berkeley Fast Filesystem:

- set of inode storage cells
- set of scattered "superblocks"
- map of disk blocks

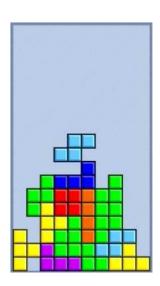


The filesystem is responsible for storing the data on the disk. So to read/write data, it needs to know in which physical blocks the actual data is located; ie how to map files to the disk blocks.

Components of the Berkeley Fast Filesystem:

- set of inode storage cells
- set of scattered "superblocks"
- map of disk blocks
- block usage summary

fstyp -v /dev/rdsk/c1t2160d0s0
| more

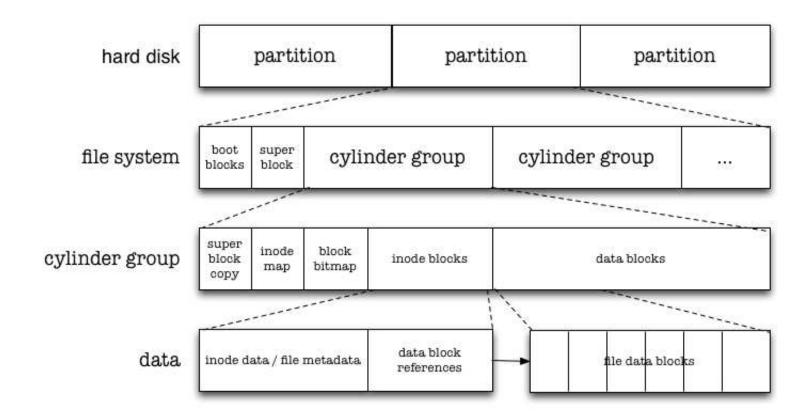


The filesystem is responsible for storing the data on the disk. So to read/write data, it needs to know in which physical blocks the actual data is located; ie how to map files to the disk blocks.

Components of the Berkeley Fast Filesystem:

- set of inode storage cells
- set of scattered "superblocks"
- map of disk blocks
- block usage summary
- set of data blocks





Information stored in an *inode*:

user owner and group owner ID's

- user owner and group owner ID's
- file type

- user owner and group owner ID's
- file type
- access mode (permissions)

- user owner and group owner ID's
- file type
- access mode (permissions)
- file access and modification time

- user owner and group owner ID's
- file type
- access mode (permissions)
- file access and modification time
- file status modification time

- user owner and group owner ID's
- file type
- access mode (permissions)
- file access and modification time
- file status modification time
- number of links to the file

- user owner and group owner ID's
- file type
- access mode (permissions)
- file access and modification time
- file status modification time
- number of links to the file
- size of the file

- user owner and group owner ID's
- file type
- access mode (permissions)
- file access and modification time
- file status modification time
- number of links to the file
- size of the file
- disk device containing this file

- user owner and group owner ID's
- file type
- access mode (permissions)
- file access and modification time
- file status modification time
- number of links to the file
- size of the file
- disk device containing this file
- \$ stat /etc/passwd

File types:

regular files

\$ stat /etc/passwd



File types:

- regular files
- directories

\$ stat /



File types:

- regular files
- directories
- special files

\$ file /dev/* | more



File types:

- regular files
- directories
- special files
- links
- \$ touch /tmp/foo
- \$ ln /tmp/foo /tmp/bar
- \$ stat /tmp/foo /tmp/bar
- \$ ln -sf /tmp/foo /tmp/bar
- \$ stat /tmp/foo /tmp/bar



File types:

- regular files
- directories
- special files
- links
- sockets

\$ stat /dev/log



File types:

- regular files
- directories
- special files
- links
- sockets
- named pipes

```
$ mkfifo /tmp/fifo
$ cat /tmp/fifo > /tmp/out &
$ stat /tmp/fifo | tee /tmp/fifo
$ cat /tmp/out
```



Homework

Repeat the examples from class. Make sure you understand the commands and how they relate to the concepts we discussed. Repeat for a different OS, for example:

- ami-b232d0db Fedora 8
- ami-23dae94a FreeBSD 10.0
- ami-65a7f30c NetBSD 6.1.2
- ami-35eb835c OmniOS 5.11

Remember to *shut down* your EC2 instances and to *delete* any unused ESB volumes!

- http://is.gd/5mndwA
- http://is.gd/ig4QP5
- http://is.gd/9YeIKh

Disk Interfaces:

- SCSI:
 - http://en.wikipedia.org/wiki/Scsi
 - scsi(4), scsictl(8);
- ATA:
 - http://www.ata-atapi.com/
 - **9**

https://en.wikipedia.org/wiki/Advanced_Technology_Attachment

https://en.wikipedia.org/wiki/Sata

Disk Interfaces:

- Serial attached SCSI:
 - https://en.wikipedia.org/wiki/Serial_attached_SCSI
- Fibre Channel:
 - https://hsi.web.cern.ch/HSI/fcs/fcs.html
 - https://en.wikipedia.org/wiki/Fibrechannel
- AoE, FCoE, iSCSI:
 - https://en.wikipedia.org/wiki/ATA_over_Ethernet
 - https://en.wikipedia.org/wiki/FCoE
 - https://en.wikipedia.org/wiki/ISCSI

Basic Disk Concepts:

- http://is.gd/3SXn5N
- disklabel(8), fdisk(8)
- format(1m)

RAID:

https://en.wikipedia.org/wiki/RAID

Basic Filesystem Concepts:

- http://is.gd/8KHnQj
- http://is.gd/YMm3JL
- newfs(8)

NFS: http://www.cis.udel.edu/~srisatha/NFS/sld001.htm