Lecture 18: File System Implementation

Instructor: Mitch Neilsen

Office: N219D

Quote of the Day

"To be a nemesis, you have to actively try to destroy something, don't you?

Really, I'm not out to destroy Microsoft.

That will just be a completely unintentional side effect. "

-- Linus Torvalds

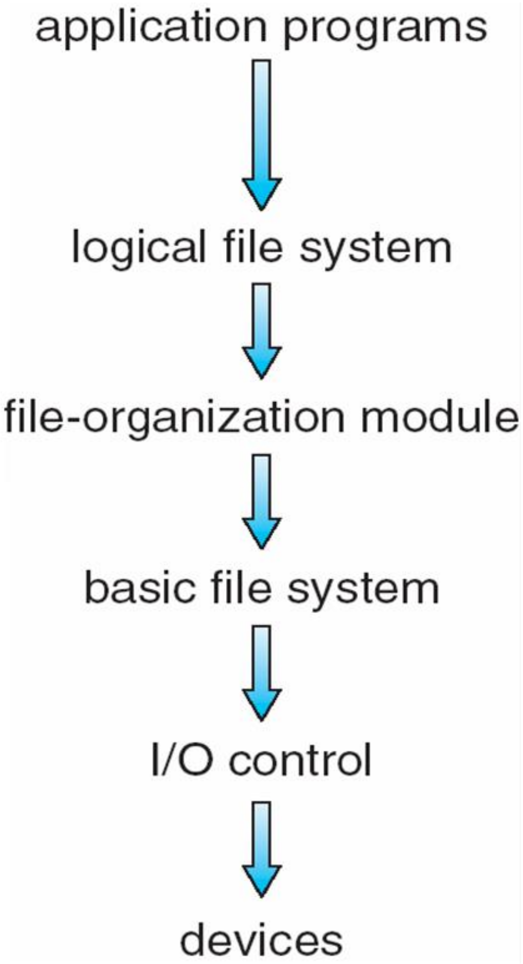
Outline – Chapters 11/12

- File System Structure
- File System Implementation
- Directory Implementation
- Allocation Methods
- Free-Space Management
- Efficiency and Performance
- Recovery
- NFS
- Example: WAFL File System

File system fun

- **File systems = the hardest part of OS**
  - More papers on file systems than any other single topic
- **Main tasks of file system:**
  - Don’t go away (ever)
  - Associate bytes with name (files)
  - Associate names with each other (directories)
  - Can implement file systems on disk, over network, in memory, in non-volatile ram (NVRAM), on tape, w/ paper.
  - We’ll focus on disk and generalize later
- **Today: files, directories, and a bit of performance**
- **Read Ch. 11/12 - File System Implementation**

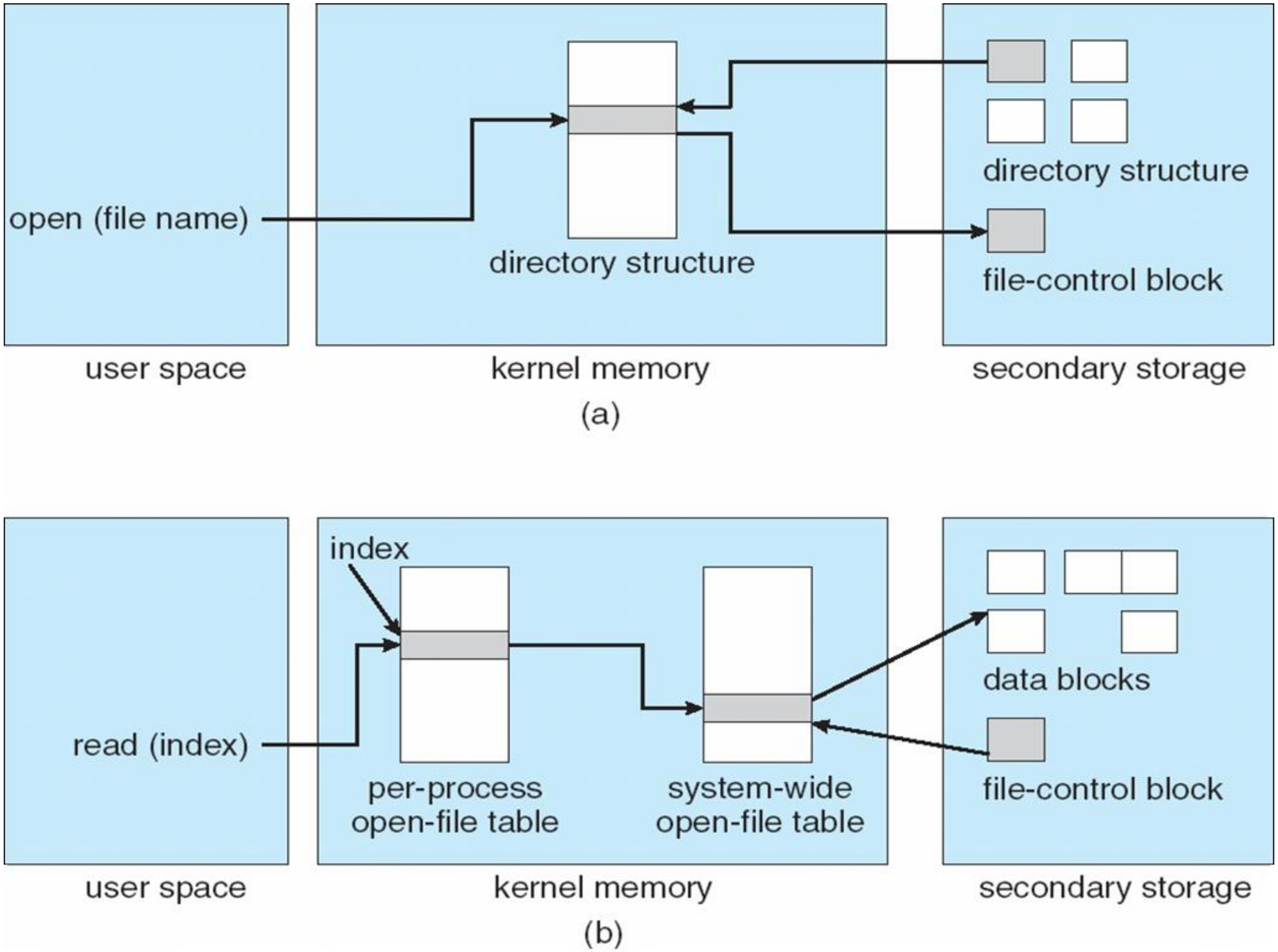
Layered File System



A Typical File Control Block

file permissions
file dates (create, access, write)
file owner, group, ACL
file size
file data blocks or pointers to file data blocks

In-Memory File System Structures



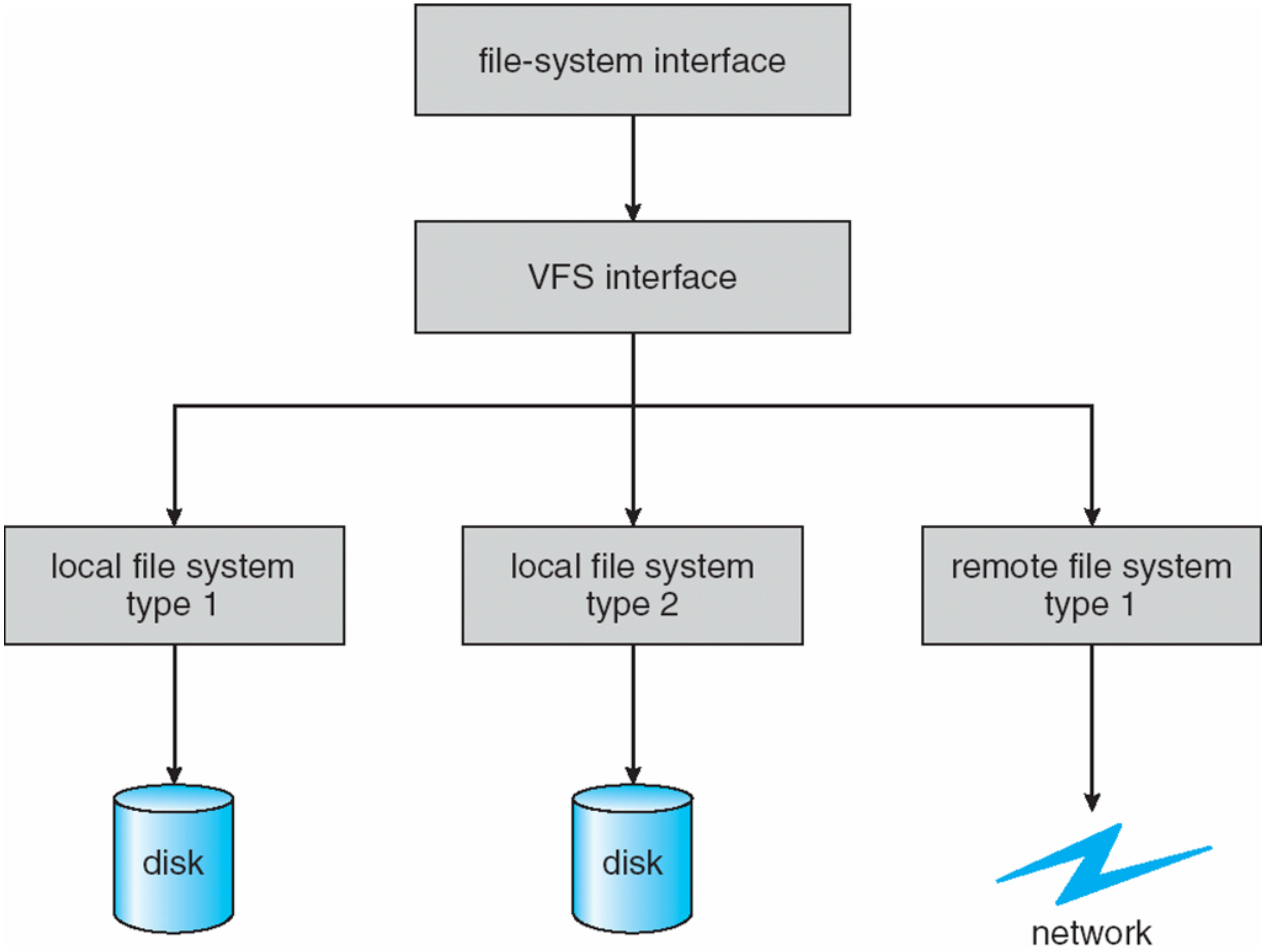
Partitions and Mounting

- Partition can be a volume containing a file system ("cooked") or **raw** – just a sequence of blocks with no file system
- Boot block can point to boot volume or boot loader set of blocks that contain enough code to know how to load the kernel from the file system
  - Or a boot management program for multi-os booting
- **Root partition** contains the OS, other partitions can hold other Oses, other file systems, or be raw
  - Mounted at boot time
  - Other partitions can mount automatically or manually
- At mount time, file system consistency checked
  - Is all metadata correct?
    - If not, fix it, try again
    - If yes, add to mount table, allow access

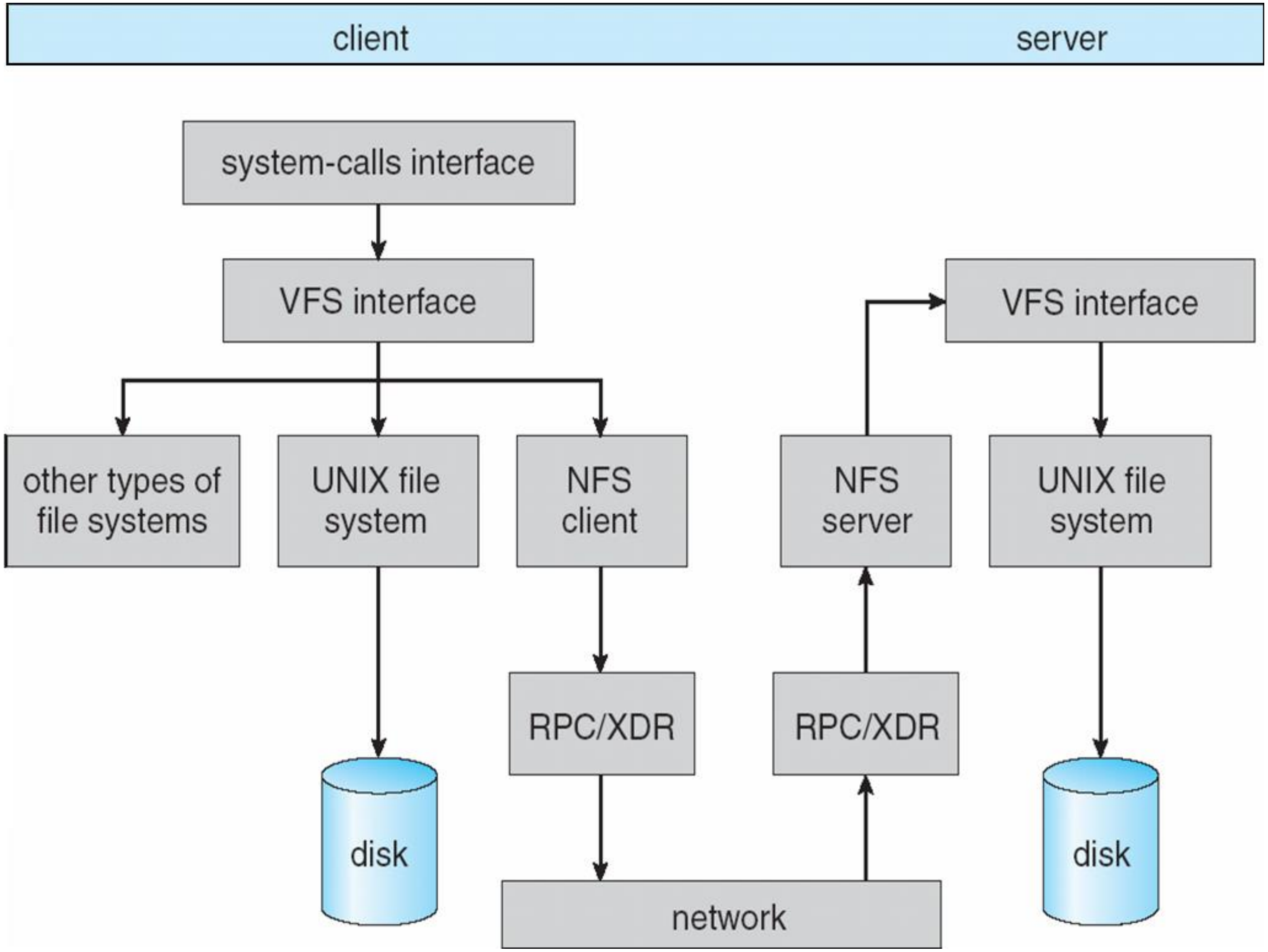
Virtual File Systems

- **Virtual File Systems (VFS)** on Unix provide an object-oriented way of implementing file systems
- VFS allows the same system call interface (the API) to be used for different types of file systems
  - Separates file-system generic operations from implementation details
  - Implementation can be one of many file systems types, or network file system
    - Implements **vnodes** which hold **inodes** or **network file details**
  - Then dispatches operation to appropriate file system implementation routines
- The API is to the VFS interface, rather than any specific type of file system

Virtual File System

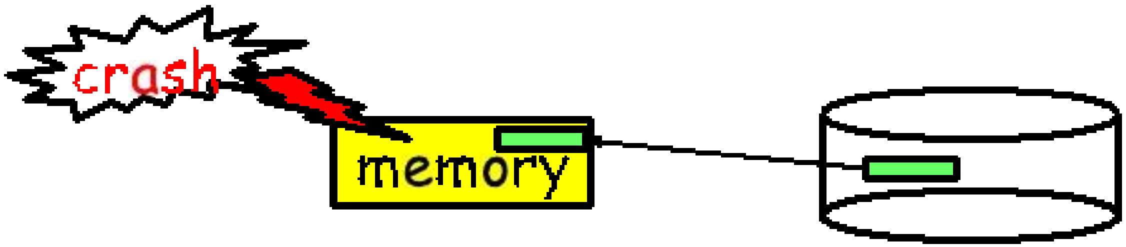


Schematic View of NFS Architecture



The medium is the message

- **Disk = First thing we've seen that doesn't go away**



  - So: Where all important state ultimately resides

- **Slow (ms access vs ns for memory)**

- **Huge (100–1,000x bigger than memory)**
  - How to organize large collection of ad hoc information?
  - Taxonomies! (Basically FS = general way to make these)

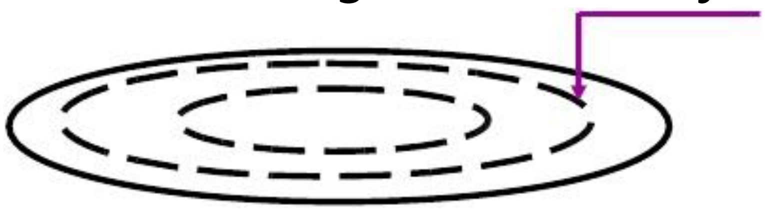
Disk vs. Memory

	Disk	MLC NAND Flash	DRAM
Smallest write	sector	sector	byte
Atomic write	sector	sector	byte/word
Random read	8 ms	75 μs	50 ns
Random write	8 ms	300 μs*	50 ns
Sequential read	100 MB/s	250 MB/s	> 1 GB/s
Sequential write	100 MB/s	170 MB/s*	> 1 GB/s
Cost	$.08–1/GB	$3/GB	$10-25/GB
Persistence	Non-volatile	Non-volatile	Volatile

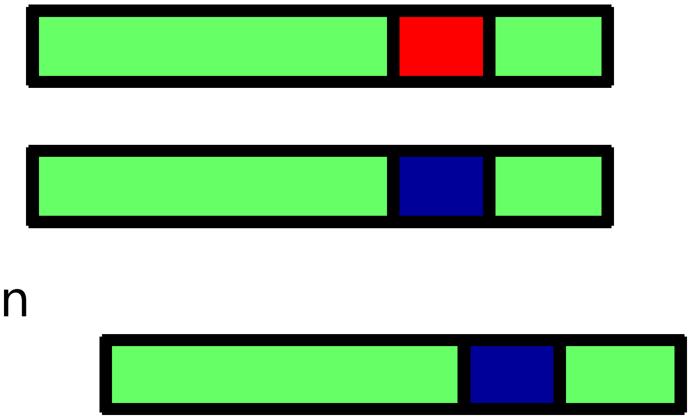
*Flash write performance degrades over time

Disk review

- **Disk reads/writes in terms of sectors, not bytes**
  - Read/write single sector or adjacent groups



- **How to write a single byte? “Read-modify-write”**
  - Read in sector containing the byte
  - Modify that byte
  - Write entire sector back to disk
  - Key: if cached, don’t need to read in

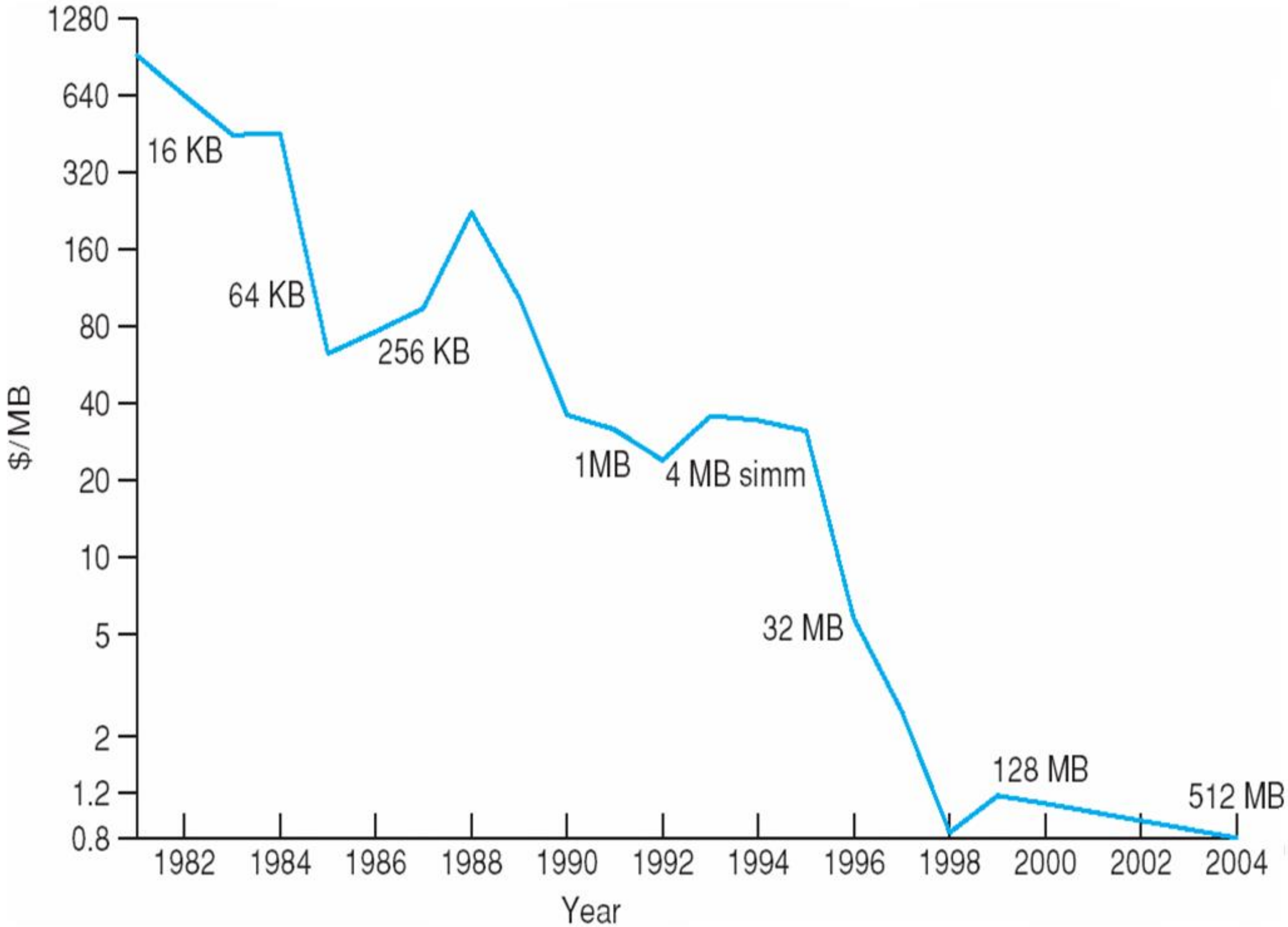


- **Sector = unit of atomicity.**
  - Sector write done completely, even if crash in middle (disk saves up enough momentum to complete)
- **Larger atomic units have to be synthesized by OS**

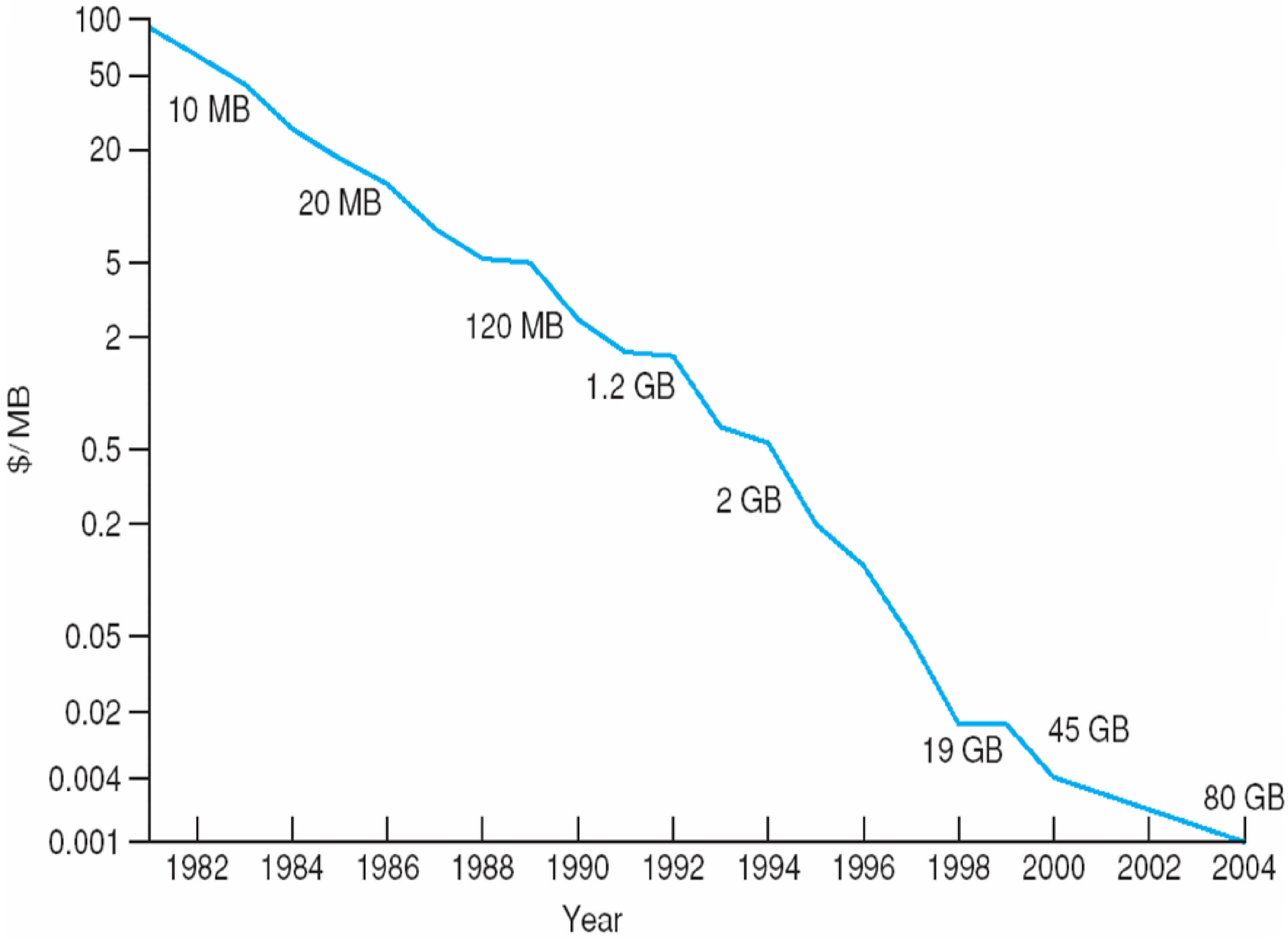
Some useful trends

- **Disk bandwidth and cost/MB improving exponentially**
  - Similar to CPU speed, memory size, etc.

Price per MB of DRAM, From 1981 to 2004



Price per MB of Magnetic Hard Disk, 1981 to 2004

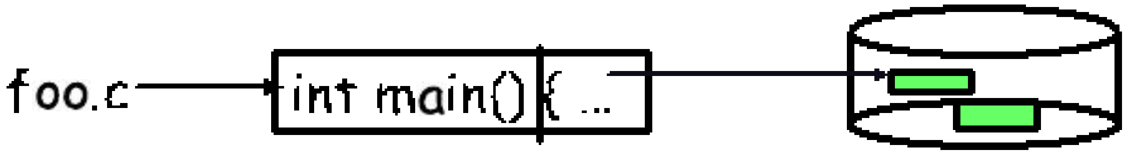


Some useful trends

- **Disk bandwidth and cost/MB improving exponentially**
  - Similar to CPU speed, memory size, etc.
- **Seek time and rotational delay improving *very* slowly**
  - Why? require moving physical object (disk arm)
- **Disk accesses a huge system bottleneck & getting worse**
  - Bandwidth increase lets system (pre-)fetch large chunks for about the same cost as small chunk.
  - Trade bandwidth for latency if you can get lots of related stuff.
  - How to get related stuff? Cluster together on disk
- **Memory size increasing faster than typical workload size**
  - More and more of workload fits in file cache
  - Disk traffic changes: mostly writes and new data

Files: named bytes on disk

- **File abstraction:**
  - **User 's view: named sequence of bytes**



  - **File system's view: collection of disk blocks**
  - **File system's job: translate name & offset to disk blocks:**

{file, offset} – – → [FS] – → disk address

- **File operations:**
  - **Create a file, delete a file**
  - **Read from file, write to file**
- **Want: operations to have as few disk accesses as possible & have minimal space overhead**

What's hard about grouping blocks?

- Like page tables, file system metadata are simply data structures used to construct mappings
  - Page table: map virtual page # to physical page #

    23 ----------→ Page table ----------→ 33

  - File meta data: map byte offset to disk block address

    418 ----------→ Unix inode ------→ 8003121

  - Directory: map name to disk address or file #

    foo.c --------→ directory ----------→ 44

FS vs. VM

- **In both settings, want location transparency**
- **In some ways, FS has easier job than VM:**
  - CPU time to do FS mappings not a big deal (= no TLB)
  - Page tables deal with sparse address spaces and random access, files often denser (0 . . . file size − 1) & ∼sequentially accessed
- **In some ways FS's problem is harder:**
  - Each layer of translation = potential disk access
  - Space a huge premium! (But disk is huge?!?!) Reason? Cache space never enough; amount of data you can get in one fetch never enough
  - Range very extreme: Many files <10 KB, some files many GB

Some working intuitions

- **FS performance dominated by # of disk accesses**
  - Each access costs ~10 milliseconds
  - Touch the disk 100 extra times = 1 *second*
  - Can easily do 100s of millions of ALU ops in same time
- **Access cost dominated by movement, not transfer:**

seek time + **rotational delay** + # bytes/disk-bw

  - Can get 50x the data for only ~3% more overhead
  - 1 sector: 10ms + 8ms + 10μs (= 512 B/(50 MB/s)) ≈ 18ms
  - 50 sectors: 10ms + 8ms + .5ms = 18.5ms
- **Observations that might be helpful:**
  - All blocks in file tend to be used together, sequentially
  - All files in a directory tend to be used together
  - All names in a directory tend to be used together

Common addressing patterns

- **Sequential:**
  - File data processed in sequential order
  - By far the most common mode
  - Example: editor writes out new file, compiler reads in file, etc
- **Random access:**
  - Address any block in file directly without passing through predecessors
  - Examples: data set for demand paging, databases
- **Keyed access**
  - Search for block with particular values
  - Examples: associative data base, index
  - Usually not provided by OS

Problem: how to track file's data

- **Disk management:**
  - Need to keep track of where file contents are on disk
  - Must be able to use this to map byte offset to disk block
  - Structure tracking a file's sectors is called an index node or ***inode***
  - File descriptors must be stored on disk, too
- **Things to keep in mind while designing file structure:**
  - Most files are small
  - Much of the disk is allocated to large files
  - Many of the I/O operations are made to large files
  - Want good sequential and good random access (what do these require?)

Allocation Methods

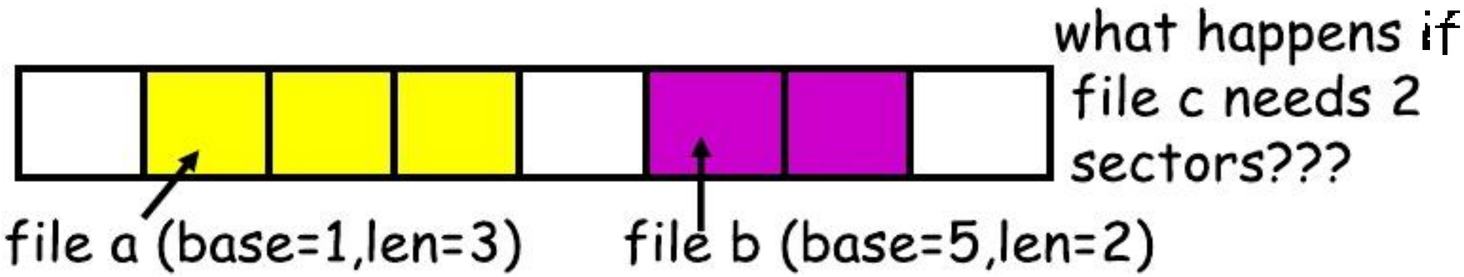
- An allocation method refers to how disk blocks are allocated for files:
- **Contiguous allocation**
- **Linked allocation**
- **Indexed allocation**

Contiguous Allocation

- Each file occupies a set of contiguous blocks on the disk
- Simple – only starting location (block #) and length (number of blocks) are required
- Random access
- Wasteful of space
- Files cannot grow

Straw man: contiguous allocation

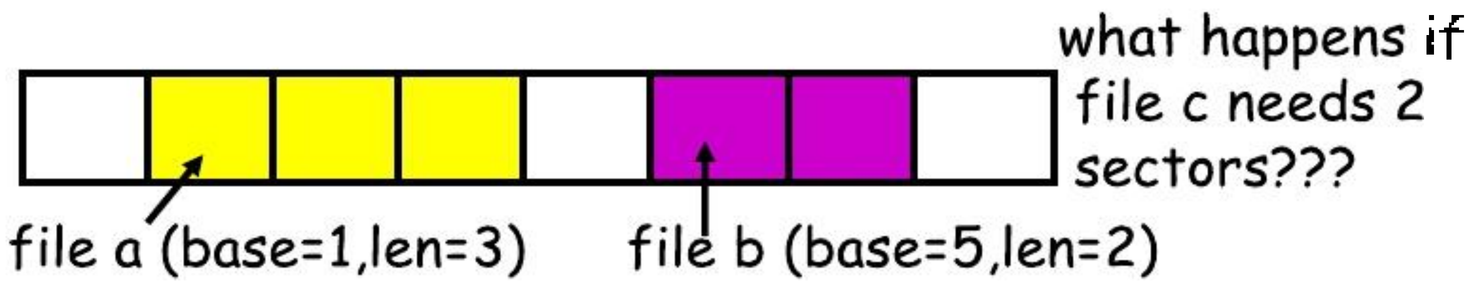
- **“Extent-based”: allocate files like segmented memory**
  - When creating a file, make the user specify pre-specify its length and allocate all space at once
  - **Inode contents: location and size**



- **Example: IBM OS/360**
- **Pros?**
- **Cons? (What VM scheme does this correspond to?)**

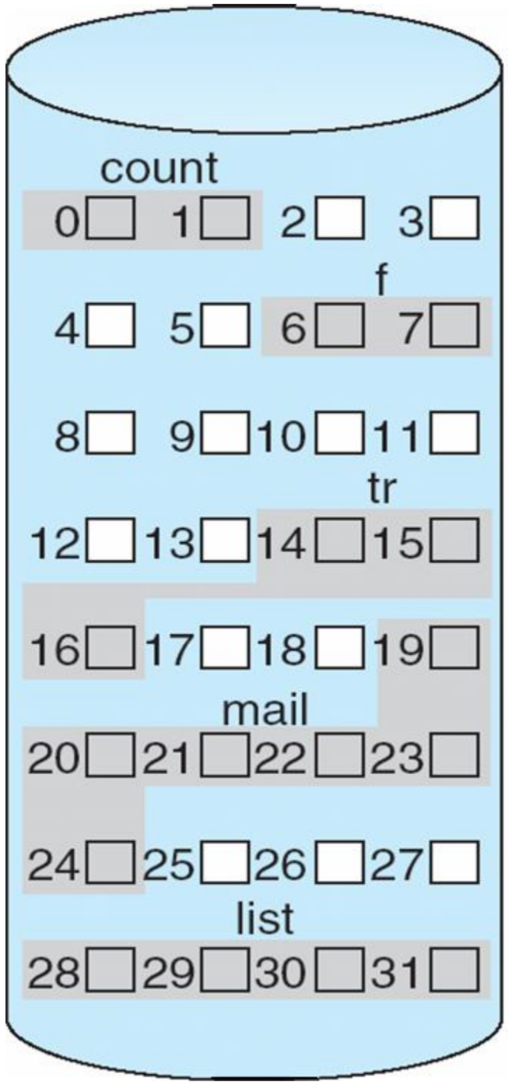
Straw man: contiguous allocation

- **"Extent-based": allocate files like segmented memory**
  - When creating a file, make the user specify pre-specify its length and allocate all space at once
  - **Inode contents: location and size**



- **Example: IBM OS/360**
- **Pros?**
  - **Simple, fast access, both sequential and random**
- **Cons? (What VM scheme does this correspond to?)**
  - **External fragmentation**

Contiguous Allocation of Disk Space



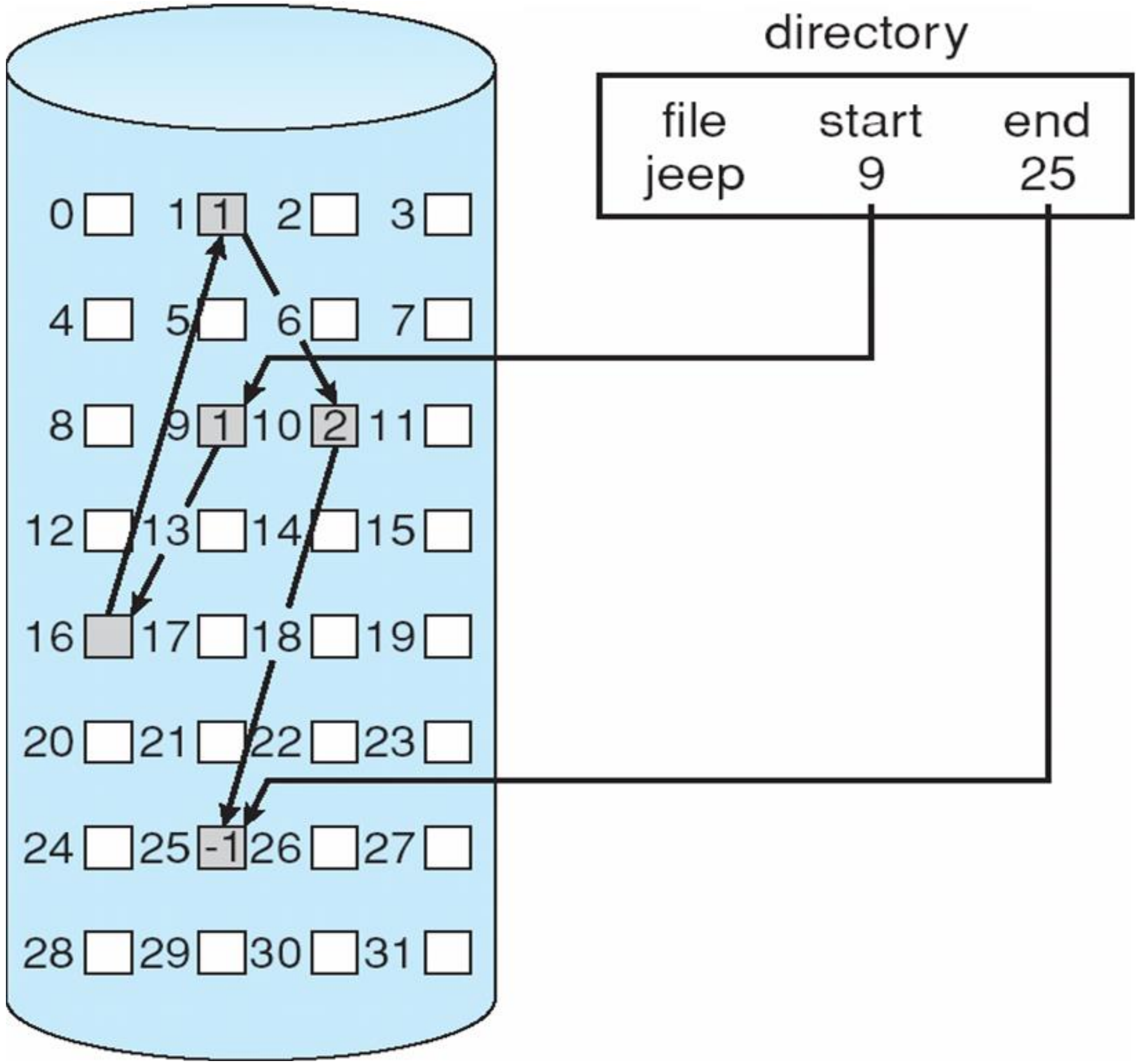
directory

file	start	length
count	0	2
tr	14	3
mail	19	6
list	28	4
f	6	2

Extent-Based Systems

- Many newer file systems (e.g., Veritas File System, Microsoft's NTFS) use a modified contiguous allocation scheme.
- Extent-based file systems allocate disk blocks in **extents** when a file is created.
- An **extent** is a contiguous block of disks
  - Extents are allocated for future file growth.
  - A file consists of one or more extents.

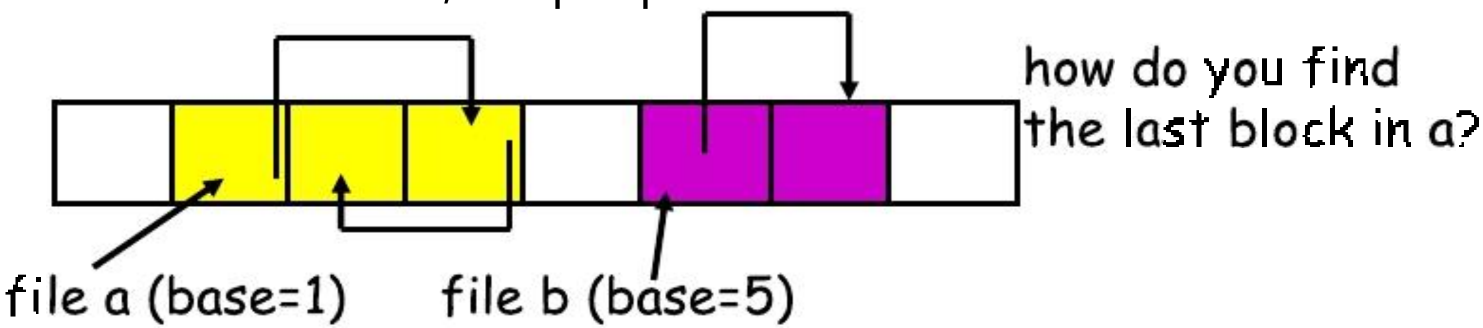
Linked Allocation



- Each file is a linked list of disk blocks: blocks may be scattered anywhere on the disk.

Linked files

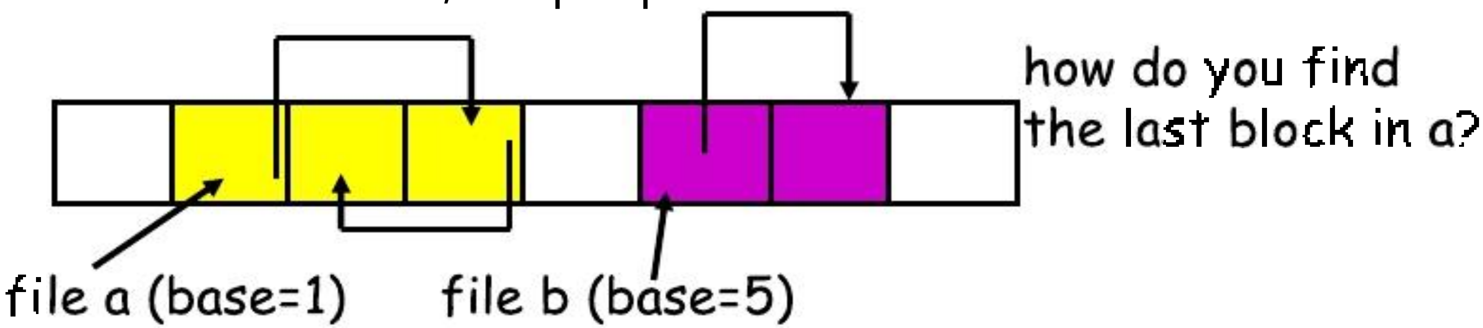
- **Basically a linked list on disk.**
  - Keep a linked list of all free blocks
  - Inode contents: a pointer to file's first block
  - In each block, keep a pointer to the next one



- **Examples (sort-of): Alto, TOPS-10, DOS FAT**
- **Pros?**
- **Cons?**

Linked files

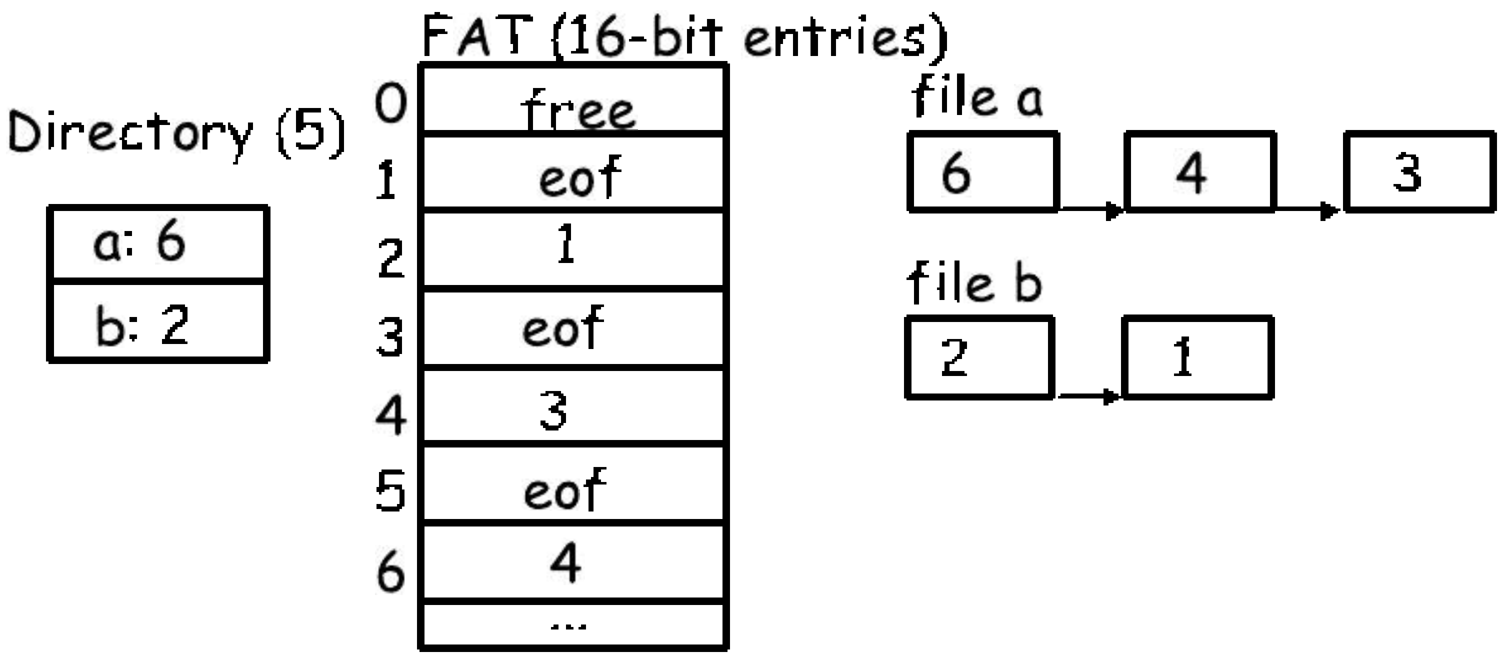
- **Basically a linked list on disk.**
  - Keep a linked list of all free blocks
  - Inode contents: a pointer to file's first block
  - In each block, keep a pointer to the next one



- **Examples (sort-of): Alto, TOPS-10, DOS FAT**
- **Pros?**
  - Easy dynamic growth & sequential access, no fragmentation
- **Cons?**
  - Linked lists on disk a bad idea because of access times
  - Pointers take up room in block, skewing alignment

Example: DOS FS (simplified)

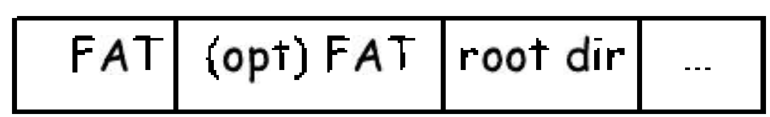
- Uses linked files. Cute: links reside in fixed-sized "file allocation table" (FAT) rather than in the blocks.



- Still do pointer chasing, but can cache entire FAT so can be cheap compared to disk access

FAT discussion

- **Entry size = 16 bits**
  - What's the maximum size of the FAT?
  - Given a 512 byte block, what's the maximum size of FS?
  - One attack: go to bigger blocks. Pros? Cons?
- **Space overhead of FAT is trivial:**
  - 2 bytes / 512 byte block = ~ 0.4% (Compare to Unix)
- **Reliability: how to protect against errors?**
  - Create duplicate copies of FAT on disk.
  - State duplication a very common theme in reliability
- **Bootstrapping: where is root directory?**
  - Fixed location on disk:

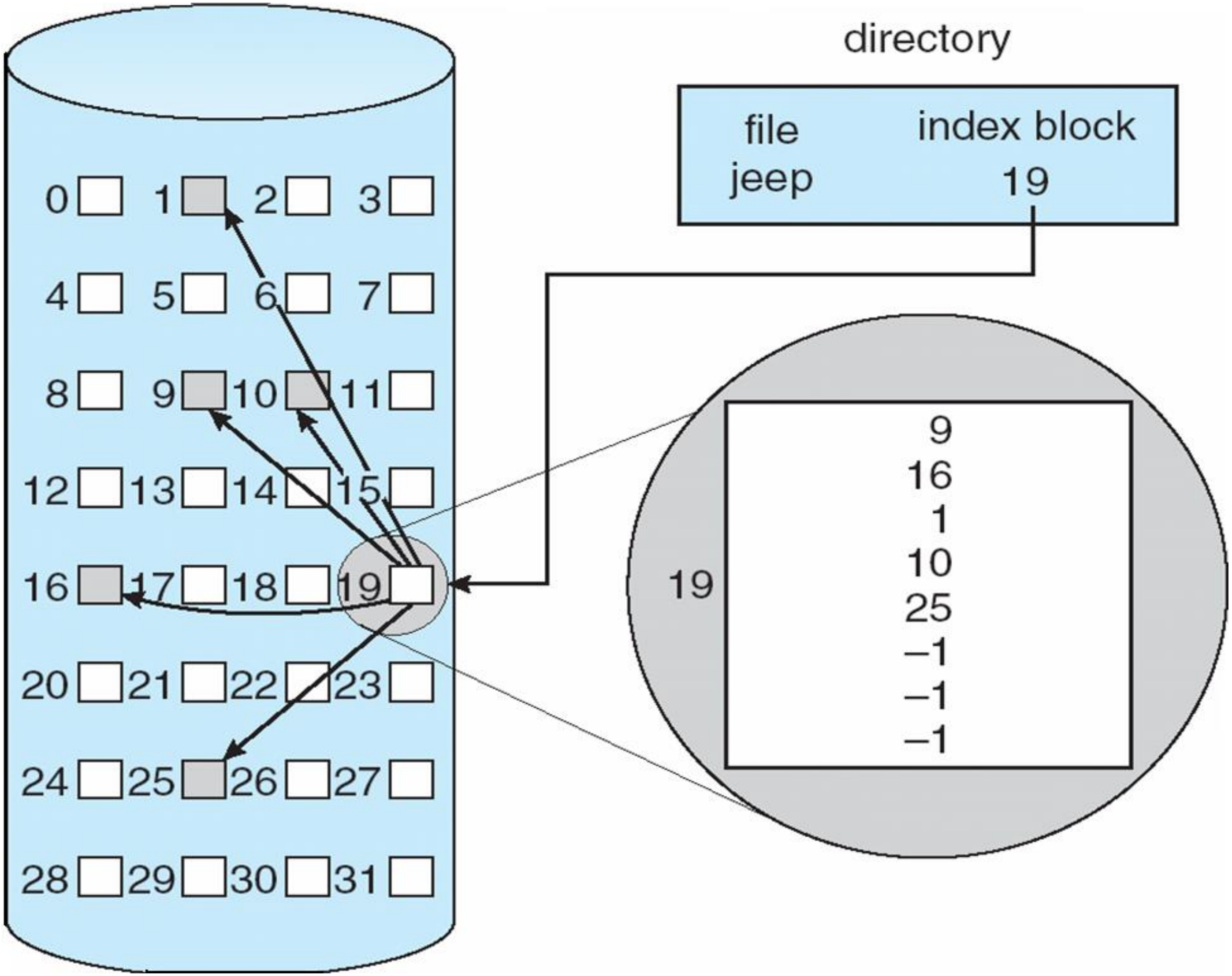


FAT discussion

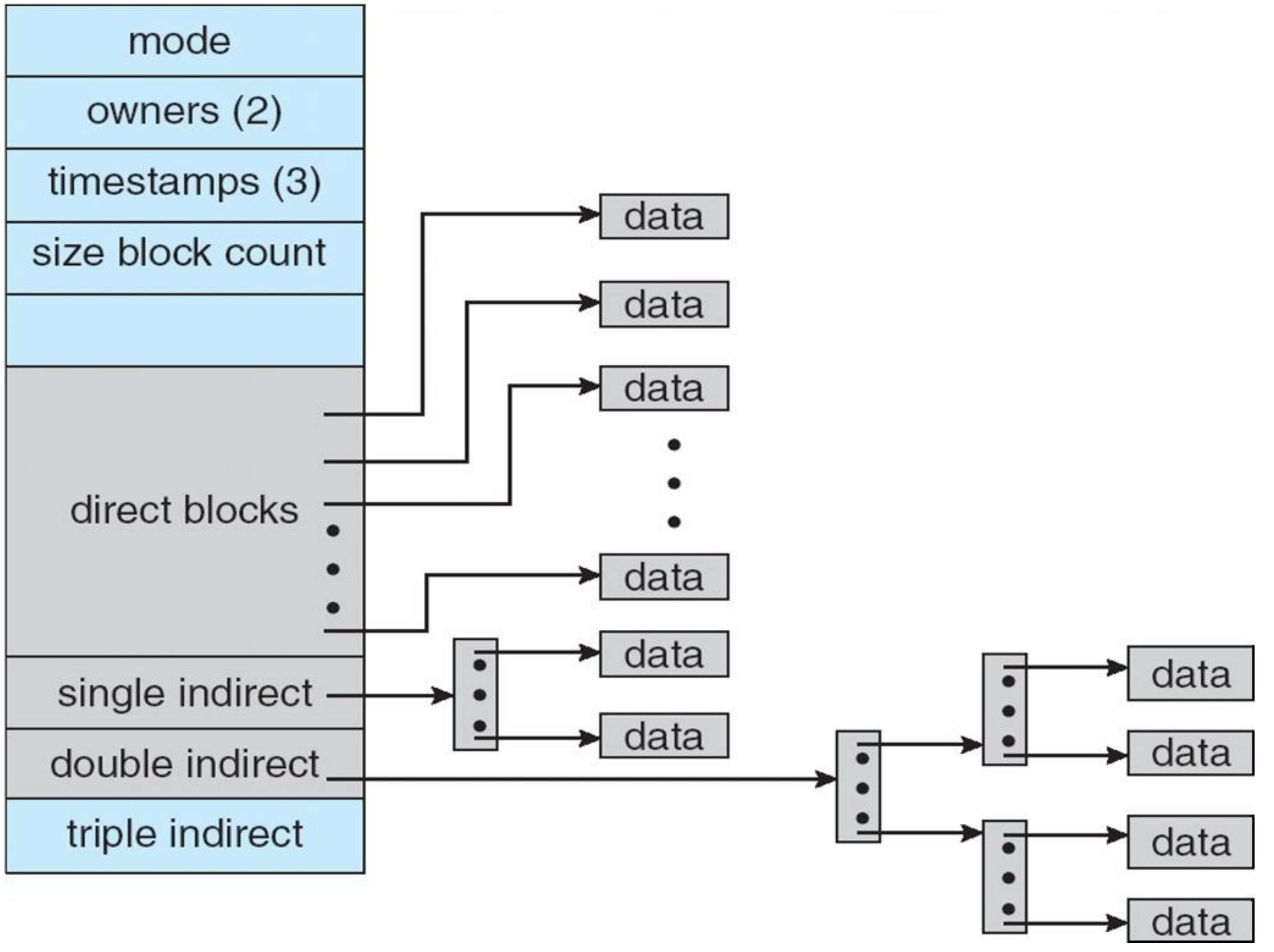
- **Entry size = 16 bits**
  - What's the maximum size of the FAT? **65,536 entries**
  - Given a 512 byte block, what's the maximum size of FS? **32 MB**
  - One attack: go to bigger blocks. Pros? Cons?
- **Space overhead of FAT is trivial:**
  - 2 bytes / 512 byte block = ~ 0.4% (Compare to Unix)
- **Reliability: how to protect against errors?**
  - Create duplicate copies of FAT on disk.
  - State duplication a very common theme in reliability
- **Bootstrapping: where is root directory?**
  - Fixed location on disk:

FAT	(opt) FAT	root dir	...

Indexed Allocation

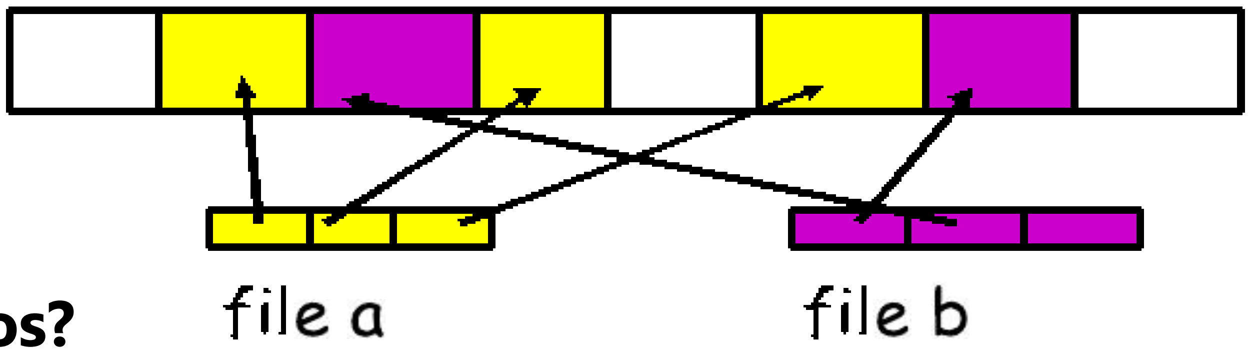


Combined Scheme: UNIX (4K bytes per block)



Indexed Allocation

- **Each file has an array holding all of it's block pointers**
  - Just like a page table, so will have similar issues
  - Max file size fixed by array's size (static or dynamic?)
  - Allocate array to hold file's block pointers on file creation
  - Allocate actual blocks on demand using free list

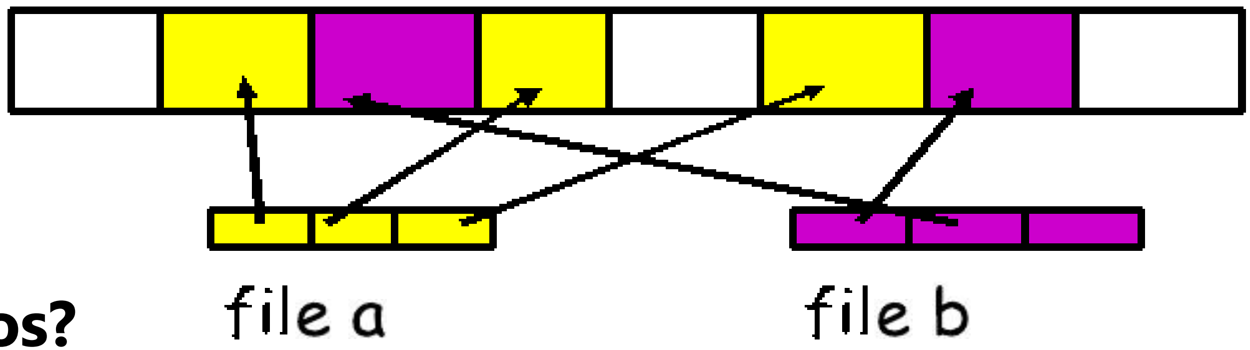


- **Pros?**

- **Cons?**

Indexed Allocation

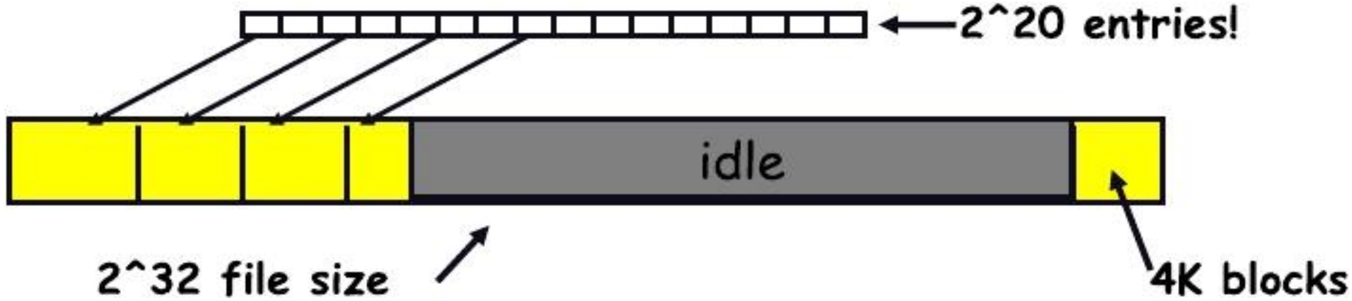
- **Each file has an array holding all of it's block pointers**
  - Just like a page table, so will have similar issues
  - Max file size fixed by array's size (static or dynamic?)
  - Allocate array to hold file's block pointers on file creation
  - Allocate actual blocks on demand using free list



- **Pros?**
  - Both sequential and random access easy
- **Cons?**
  - Mapping table requires large chunk of contiguous space
    . . . Same problem we were trying to solve initially

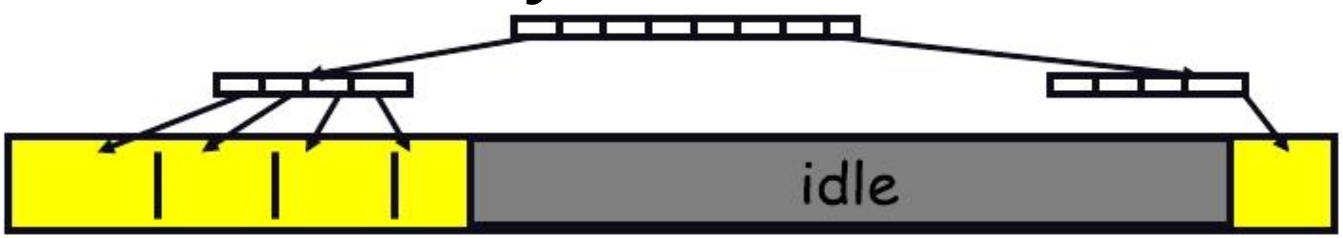
Indexed Allocation

- **Issues same as in page tables**



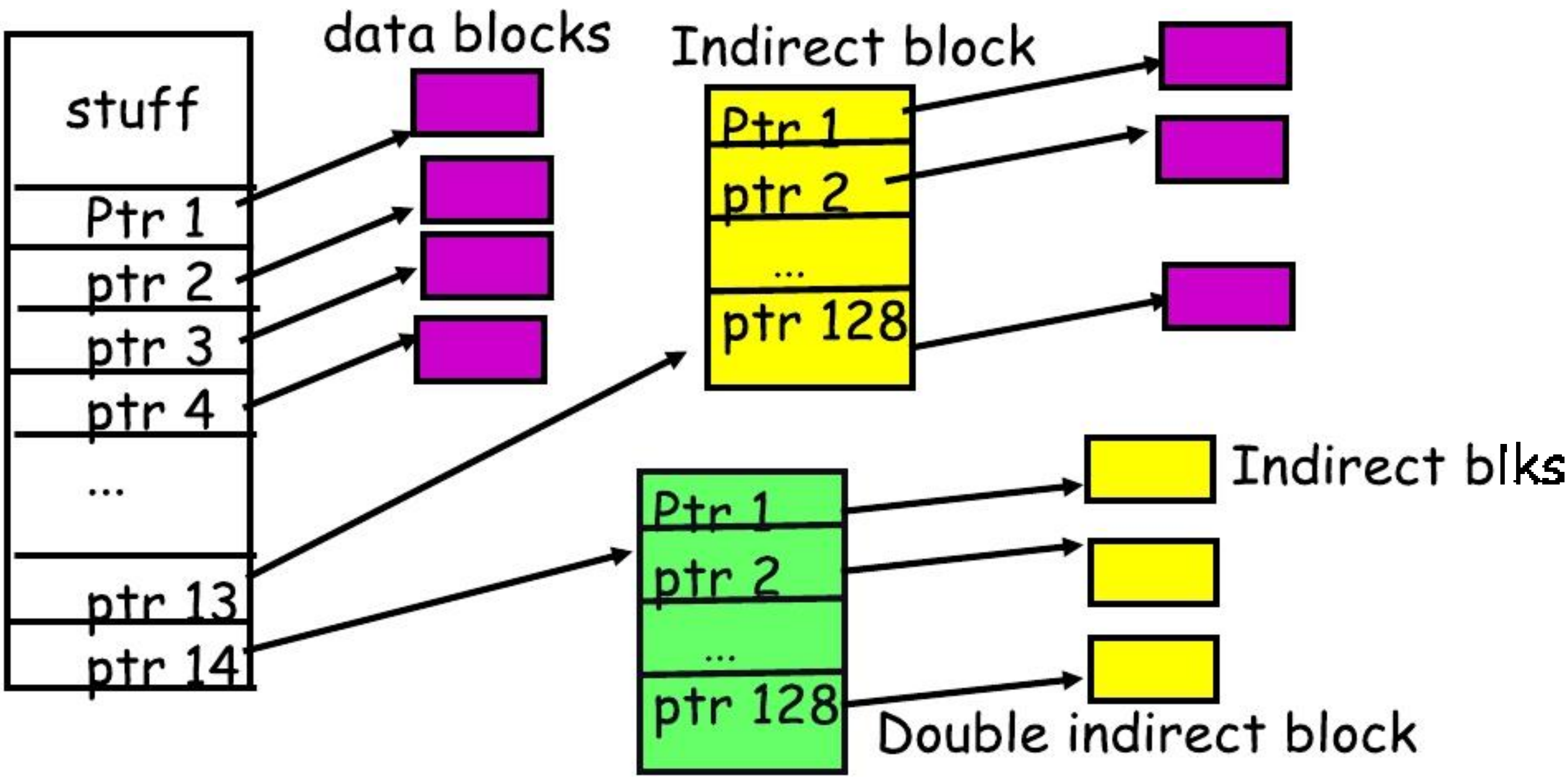
  - Large possible file size = lots of unused entries
  - Large actual size? table needs large contiguous disk chunk

- **Solve identically: small regions with index array, this array with another array, . . . Downside?**



Multi-level indexed files (old BSD FS)

- inode = 14 block pointers + "stuff"

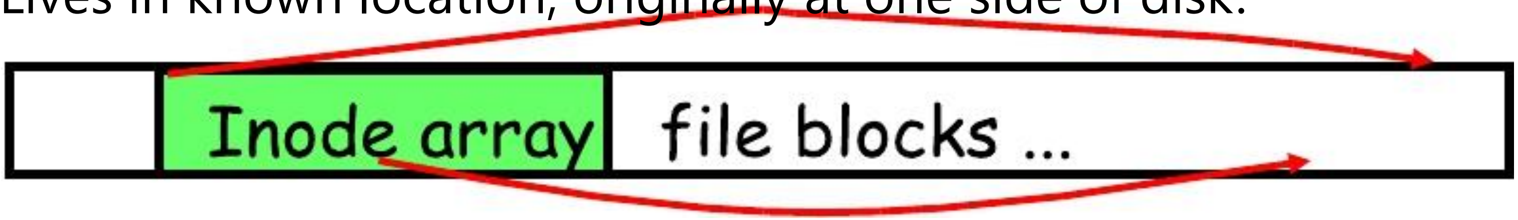


Old BSD FS discussion

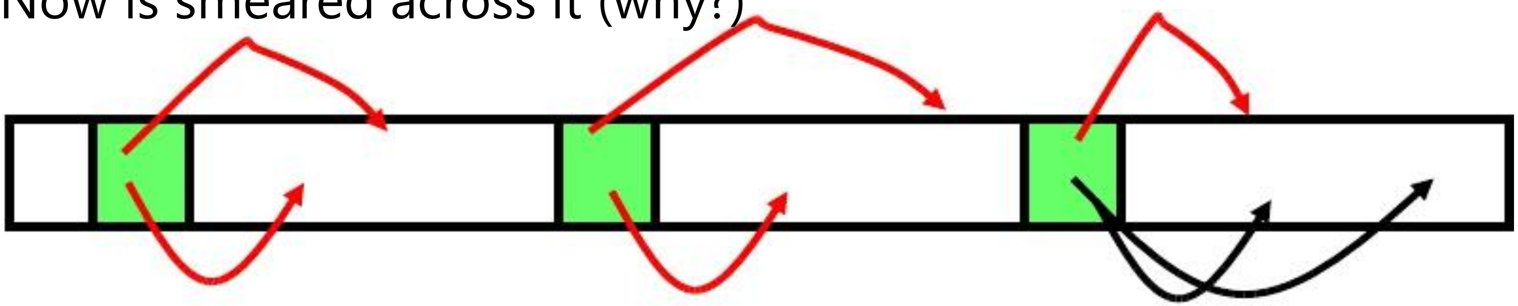
- **Pros:**
  - Simple, easy to build, fast access to small files
  - Maximum file length fixed, but large.
- **Cons:**
  - What is the worst case # of accesses?
  - What is the worst-case space overhead? (e.g., 13 block file)
- **An empirical problem:**
  - Because you allocate blocks by taking them off unordered freelist, meta data and data get strewn across disk

More about inodes

- **Inodes are stored in a fixed-size array**
  - Size of array fixed when disk is initialized; can't be changed
  - Lives in known location, originally at one side of disk:



  - Now is smeared across it (why?)



  - The index of an inode in the inode array called an i-number
  - Internally, the OS refers to files by i-number
  - When file is opened, inode brought in memory
  - Written back when modified and file closed or time elapses

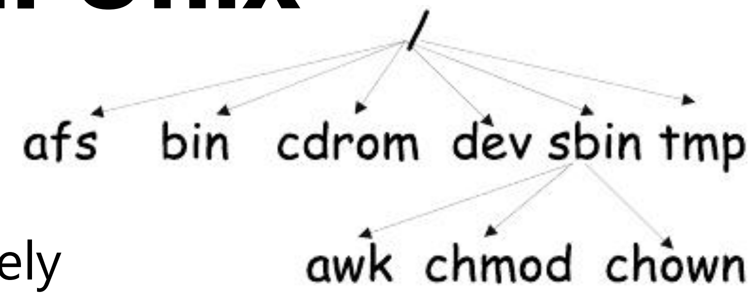
Directories

- **Problem:**
  - "Spend all day generating data, come back the next morning, want to use it." F. Corbato, on why files/dirs invented.
- **Approach 0: Have users remember where on disk their files are**
  - (E.g., like remembering your social security or bank account #)
- **Yuck. People want human digestible names**
  - We use directories to map names to file blocks
- **Next: What is in a directory and why?**

A short history of directories

- **Approach 1: Single directory for entire system**
  - Put directory at known location on disk
  - Directory contains name, i-number pairs
  - If one user uses a name, no one else can
  - Many ancient personal computers work this way
- **Approach 2: Single directory for each user**
  - Still clumsy, and ls on 10,000 files is a real pain
- **Approach 3: Hierarchical name spaces**
  - Allow directory to map names to files *or other dirs*
  - File system forms a tree (or graph, if links allowed)
  - Large name spaces tend to be hierarchical (ip addresses, domain names, scoping in programming languages, etc.)

Hierarchical Unix



- **Used since CTSS (1960s)**
  - Unix picked up and used really nicely
- **Directories stored on disk just like regular files**
  - Inode contains special flag bit set
  - User 's can read just like any other file
  - Only special programs can write (why?)
  - Inodes at fixed disk location
  - File pointed to by the index may be another directory
  - Makes FS into hierarchical tree (what needed to make a DAG?)

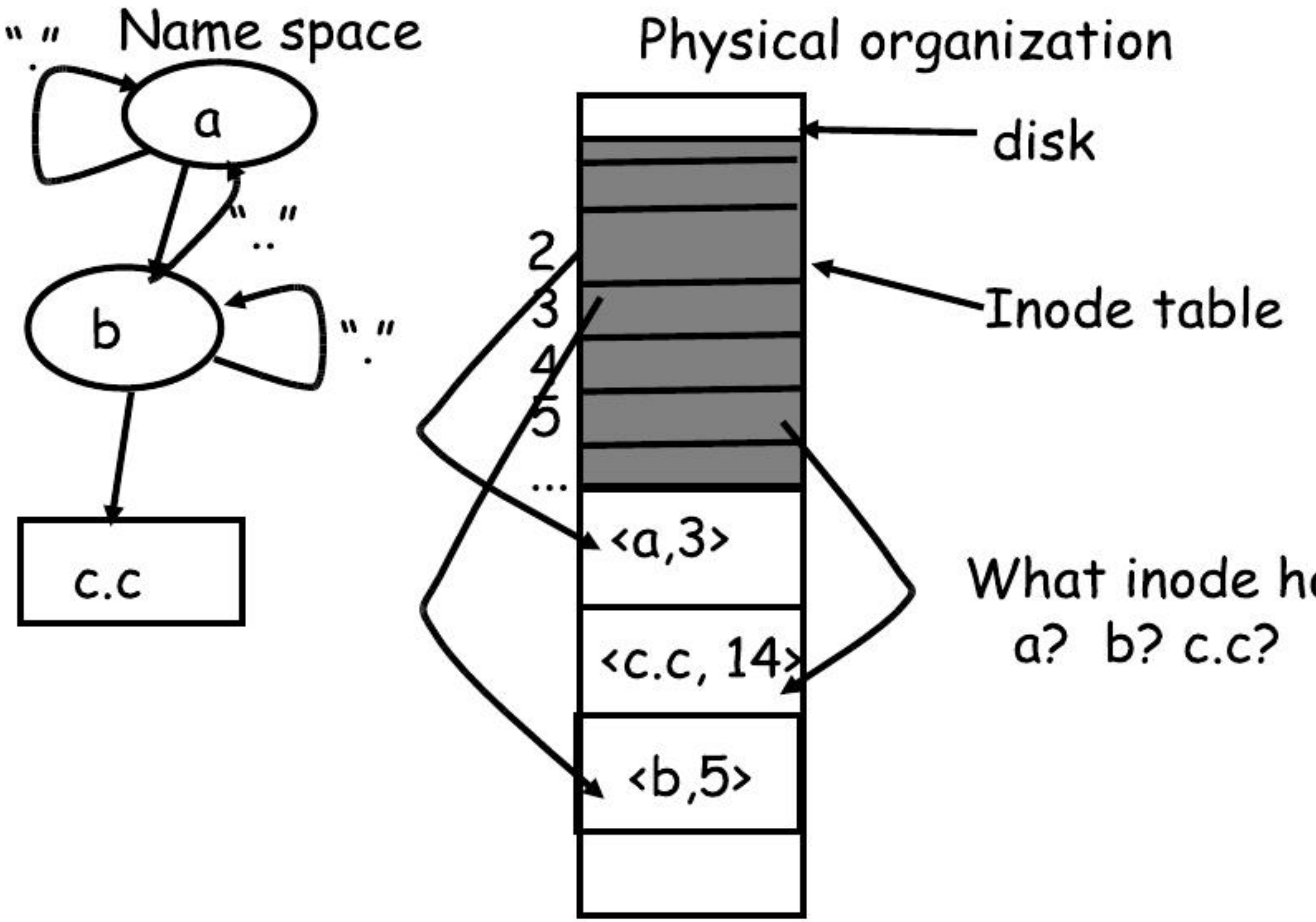
<name,inode#>
<afs,1021>
<tmp,1020>
<bin,1022>
<cdrom,4123>
<dev,1001>
<sbin,1011>
⋮

- **Simple, plus speeding up file ops speeds up dir ops!**

Naming magic

- **Bootstrapping: Where do you start looking?**
  - Root directory always inode #2 (0 and 1 historically reserved)
- **Special names:**
  - Root directory: “/”
  - Current directory: “.”
  - Parent directory: “..”
- **Special names not implemented in FS:**
  - User ’s home directory: “~”
  - Globbing: “foo.*” expands to all files starting “foo.”
- **Using the given names, only need two operations to navigate the entire name space:**
  - cd *name*: move into (change context to) directory *name*
  - ls : enumerate all names in current directory (context)

Unix example: /a/b/c.c



What inode holds file for a? b? c.c?

Default context: working directory

- **Cumbersome to constantly specify full path names**
  - In Unix, each process associated with a "current working directory"
  - File names that do not begin with "/" are assumed to be relative to the working directory, otherwise translation happens as before
- **Shells track a default list of active contexts**
  - A "search path" for programs you run
  - Given a search path $A : B : C$, a shell will check in A, then check in B, then check in C
  - Can escape using explicit paths: "./foo"
- **Example of locality**

Hard and soft links (synonyms)

- **More than one dir entry can refer to a given file**
  - Unix stores count of pointers ("hard links") to inode
  - To make: "ln foo bar" creates a synonym (bar) for *file* foo

foo bar

inode #31279
refcount = 2

- **Soft links = synonyms for *names***
  - Point to a file (or dir) *name*, but object can be deleted from underneath it (or never even exist).
  - Unix implements like directories: inode has special "sym link" bit set and contains pointed to name

    "ln -s /bar baz"

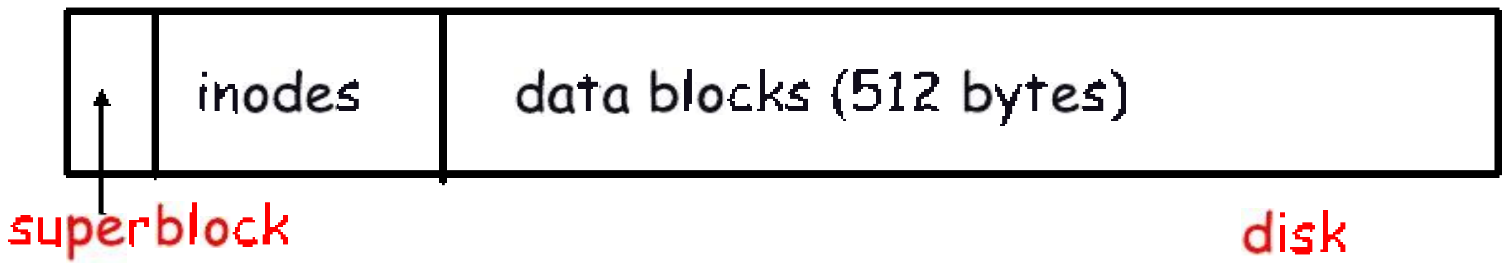
baz

"/bar"
refcount=1

  - When the file system encounters a symbolic link it automatically translates it (if possible).

Case study: speeding up FS

- **Original Unix FS: Simple and elegant:**



- **Components:**
  - Data blocks
  - Inodes (directories represented as files)
  - Hard links
  - Superblock. (specifies number of blks in FS, counts of max # of files, pointer to head of free list)
- **Problem: slow**
  - Only gets 20Kb/sec (2% of disk maximum) even for sequential disk transfers!

A plethora of performance costs

- **Blocks too small (512 bytes)**
  - File index too large
  - Too many layers of mapping indirection
  - Transfer rate low (get one block at time)
- **Bad clustering of related objects:**
  - Consecutive file blocks not close together
  - Inodes far from data blocks
  - Inodes for directory not close together
  - Poor enumeration performance: e.g., “ls”, “grep foo *.c”
- **Next: how FFS fixes these problems (to a degree)**

FFS = Berkeley Fast File System = Unix File System = UFS

Problem: Internal fragmentation

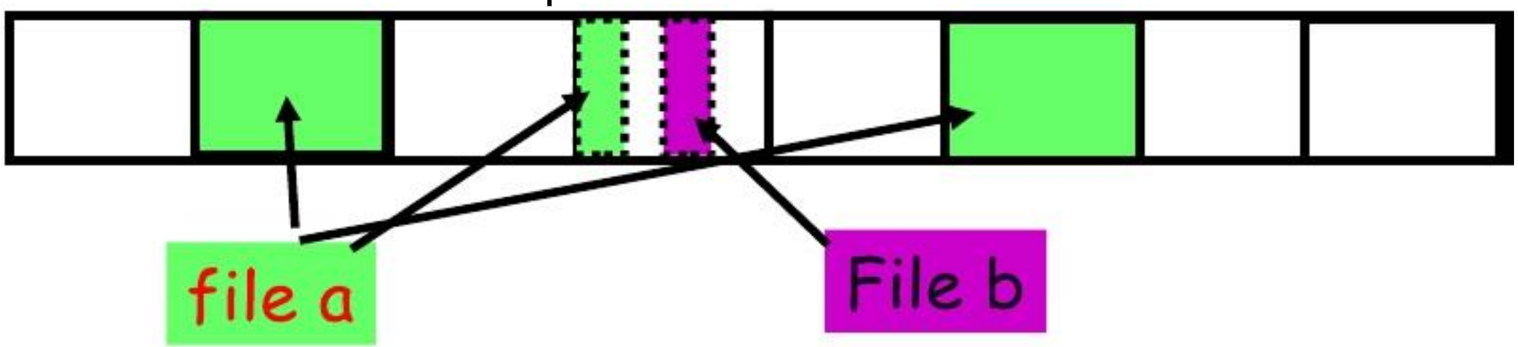
- **Block size was to small in original Unix FS**
- **Why not just make bigger?**

Block size	space wasted	file bandwidth
512	6.9%	2.6%
1024	11.8%	3.3%
2048	22.4%	6.4%
4096	45.6%	12.0%
1MB	99.0%	97.2%

- **Bigger block increases bandwidth, but how to deal with wastage (“internal fragmentation”)?**
  - Use idea from malloc: split unused portion.

Solution: fragments

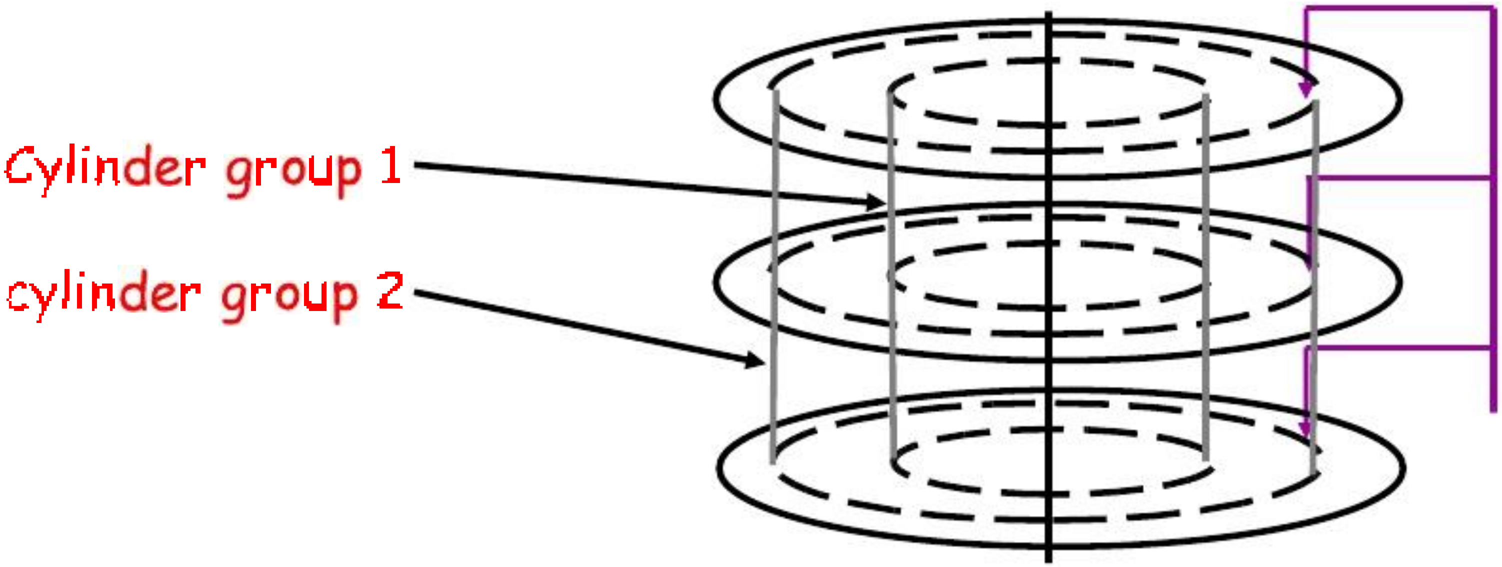
- **BSD FFS = UFS:**
  - Has large block size (4096 or 8192)
  - Allow large blocks to be chopped into small ones (“fragments”)
  - Used for little files and pieces at the ends of files



- **Best way to eliminate internal fragmentation?**
  - Variable sized splits of course
  - Why does FFS use fixed-sized fragments (1024, 2048)?

Clustering related objects in FFS

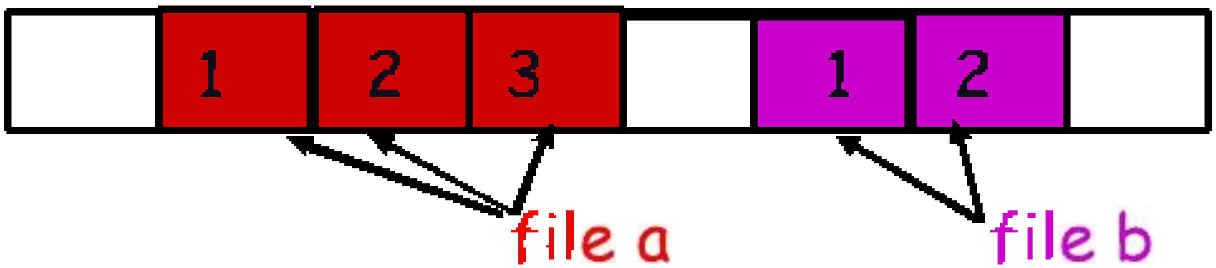
- **Group 1 or more consecutive cylinders into a "*cylinder group*"**



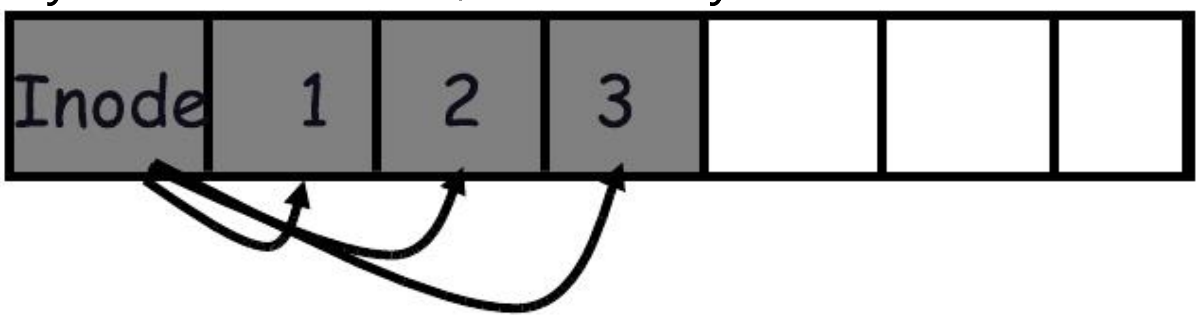
  - Key: can access any block in a cylinder without performing a seek. Next fastest place is adjacent cylinder.
  - Tries to put everything related in same cylinder group
  - Tries to put everything not related in different group (?!)

Clustering in FFS

- **Tries to put sequential blocks in adjacent sectors**
  - (Access one block, probably access next)



- **Tries to keep inode in same cylinder as file data:**
  - (If you look at inode, most likely will look at data too)



- **Tries to keep all inodes in a dir in same cylinder group**
  - Access one name, frequently access many, e.g., "ls -l"

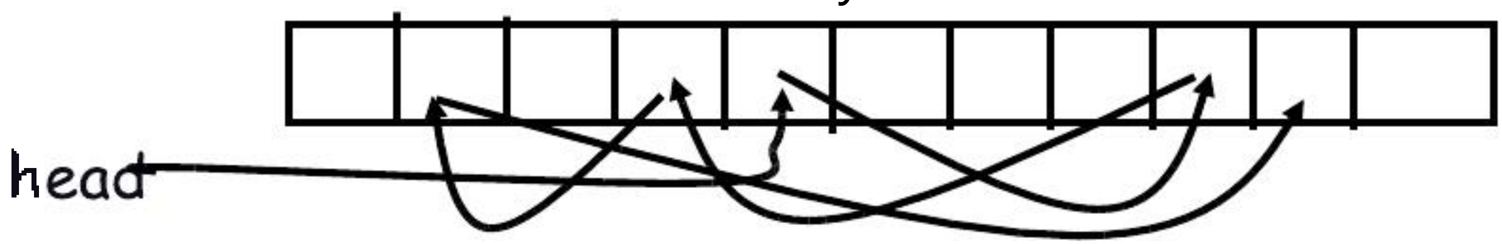
What does a cylinder group look like?

- **Basically a mini-Unix file system:**

- **How to ensure there's space for related stuff?**
  - Place different directories in different cylinder groups
  - Keep a "free space reserve" so can allocate near existing things
  - When file grows too big (1MB) send its remainder to different cylinder group.

Finding space for related objects

- **Old Unix (& dos): Linked list of free blocks**
  - Just take a block off of the head. Easy.



  - Bad: free list gets jumbled over time. Finding adjacent blocks hard and slow
- **FFS: switch to bit-map of free blocks**
  - 1010101111111000001111111000101100
  - Easier to find contiguous blocks.
  - Small, so usually keep entire thing in memory
  - Key: keep a reserve of free blocks. Makes finding a close block easier

Using a bitmap

- **Usually keep entire bitmap in memory:**
  - 4G disk / 4K byte blocks. How big is map?
- **Allocate block close to block x?**
  - Check for blocks near bmap[x/32]
  - If disk almost empty, will likely find one near
  - As disk becomes full, search becomes more expensive and less effective.
- **Trade space for time (search time, file access time)**
- **Keep a reserve (e.g., 10%) of disk always free, ideally scattered across disk**
  - Don’t tell users (df → 110% full)
  - With 10% free, can almost always find one of them free

So what did we gain?

- **Performance improvements:**
  - Able to get 20-40% of disk bandwidth for large files
  - 10-20x original Unix file system!
  - Better small file performance (why?)
- **Is this the best we can do? No.**
- **Block based rather than extent based**
  - Name contiguous blocks with single pointer and length
  - (Linux ext2fs)
- **Writes of metadata done synchronously**
  - Really hurts small file performance
  - Make asynchronous with write-ordering ("soft updates") or logging (the episode file system, ~LFS)
  - Play with semantics (/tmp file systems)

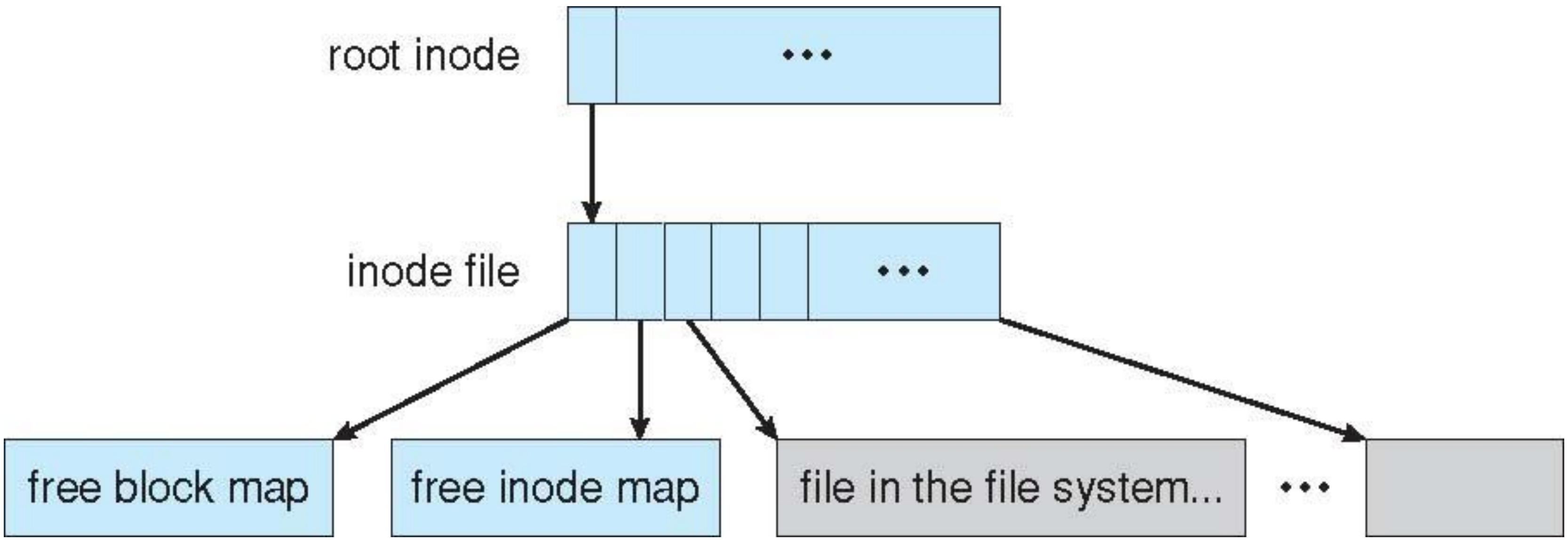
Other hacks

- **Obvious:**
  - Big file cache.
- **Fact: no rotation delay if get whole track.**
  - How to use?
- **Fact: transfer cost negligible.**
  - Recall: Can get 50x the data for only ~3% more overhead
  - 1 sector: 10ms + 8ms + 10μs (= 512 B/(50 MB/s)) ≈ 18ms
  - 50 sectors: 10ms + 8ms + .5ms = 18.5ms
  - How to use?
- **Fact: if transfer huge, seek + rotation negligible**
  - Hoard data, write out MB at a time.

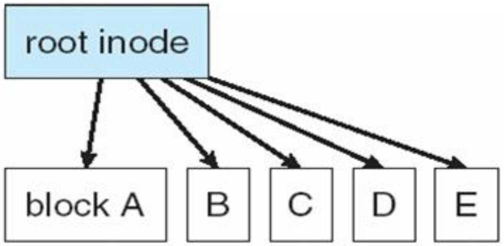
Example: WAFL File System

- Used on Network Appliance “Filers” – distributed file system appliances
- “Write-anywhere file layout”
- Serves up NFS, CIFS, http, ftp
- Random I/O optimized, write optimized
  - NVRAM for write caching
- Similar to Berkeley Fast File System, with extensive modifications

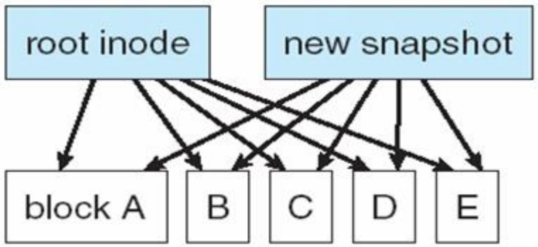
The WAFL File Layout



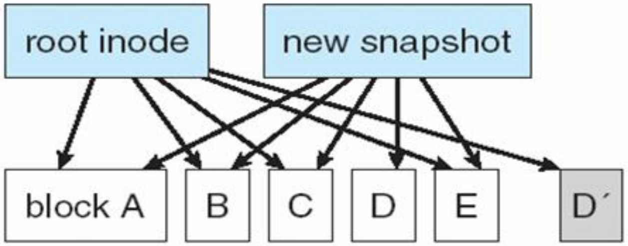
Snapshots in WAFL



(a) Before a snapshot.



(b) After a snapshot, before any blocks change.



(c) After block D has changed to D´.

Summary

- Read Ch. 1-12
- Processes and Threads (Ch. 4)
- Process Scheduling (Ch. 5)
- Synchronization (Ch. 6)
- Deadlock (Ch. 7)
- Memory Management (Ch. 8)
- Virtual Memory (Ch. 9)
- Mass-Storage Structure (Ch. 10)
- File System Interface (Ch. 11)
- File System Implementation (Ch. 12)
- Project #2 – System Calls and User-Level Processes