Advanced Programming in the UNIX Environment

Week 02, Segment 4: File Sharing

Department of Computer Science
Stevens Institute of Technology

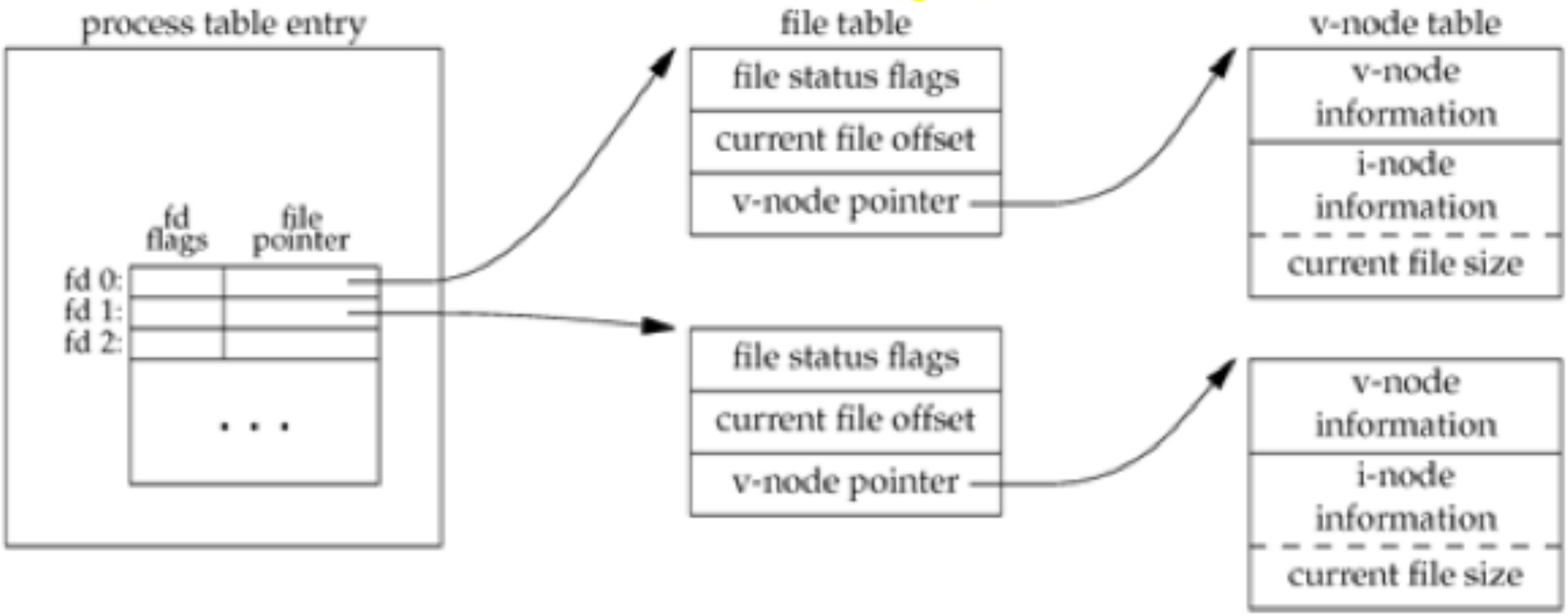
Jan Schaumann
jschauma@stevens.edu
https://stevens.netmeister.org/631/

File Sharing

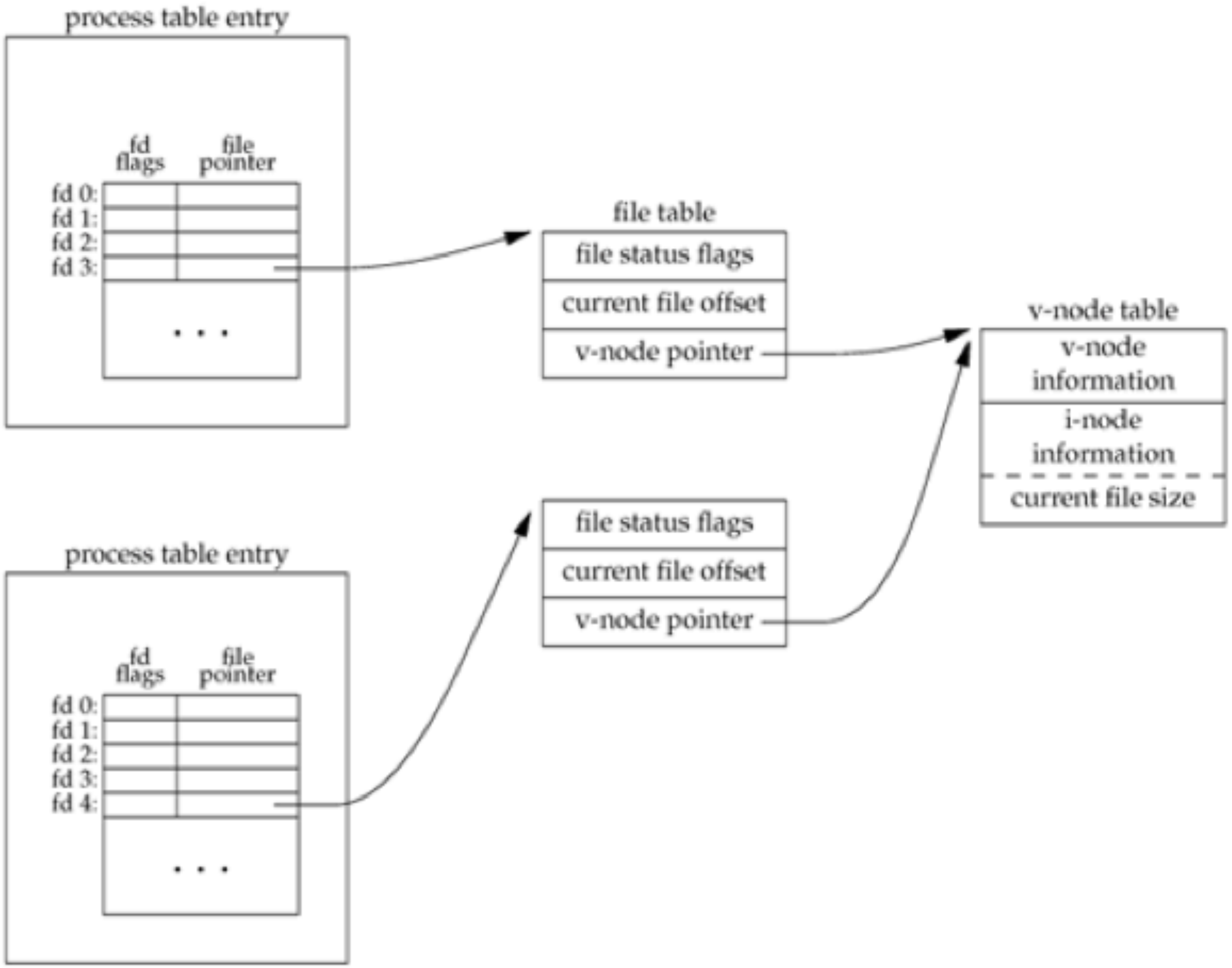
Since UNIX is a multi-user/multi-tasking system, it is conceivable (and useful) if more than one process can act on a single file simultaneously. In order to understand how this is accomplished, we need to examine some kernel data structures which relate to files.

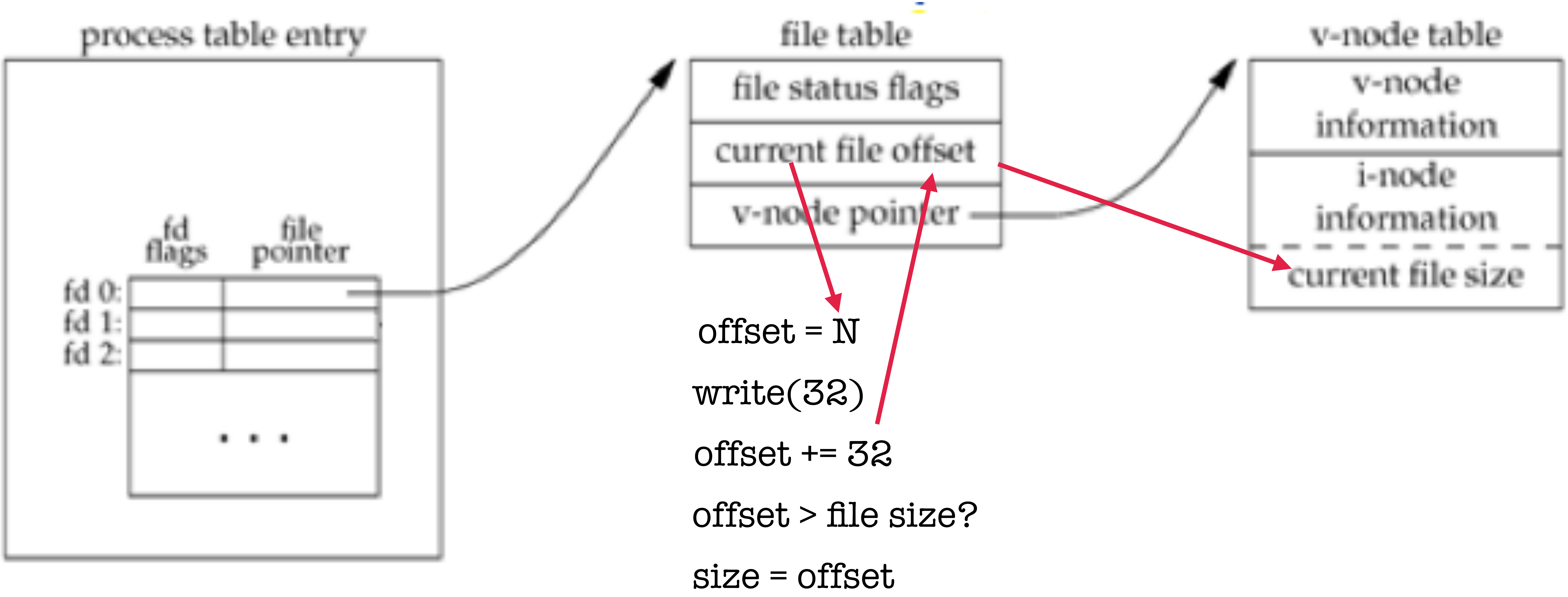
- each *process table entry* has a table of file descriptors, which contain:
    - the file descriptor flags (e.g. FD_CLOEXEC, see fcntl(2))
    - a pointer to a file table entry
- the kernel maintains a *file table*; each entry contains
    - file status flags (O_APPEND, O_SYNC, O_RDONLY, etc.)
    - current offset
    - pointer to a vnode table entry
- a *vnode* structure contains
    - vnode information
    - inode information (such as current file size)

File Sharing



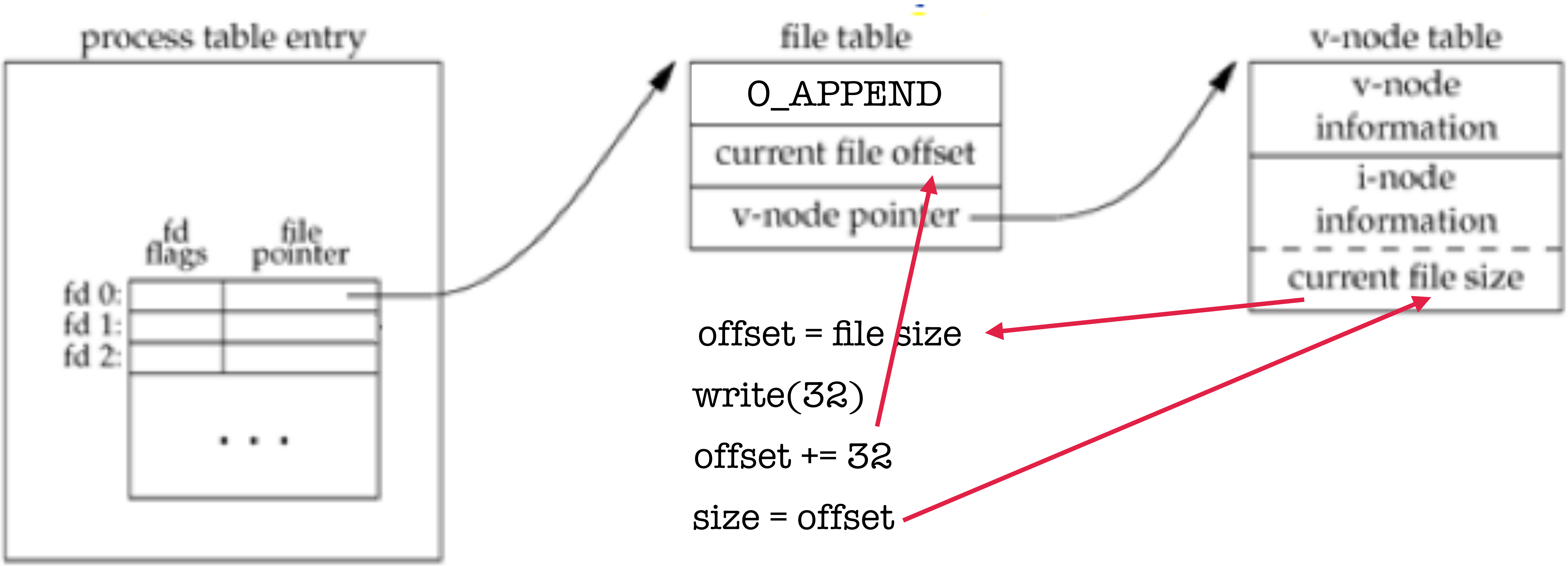
File Sharing





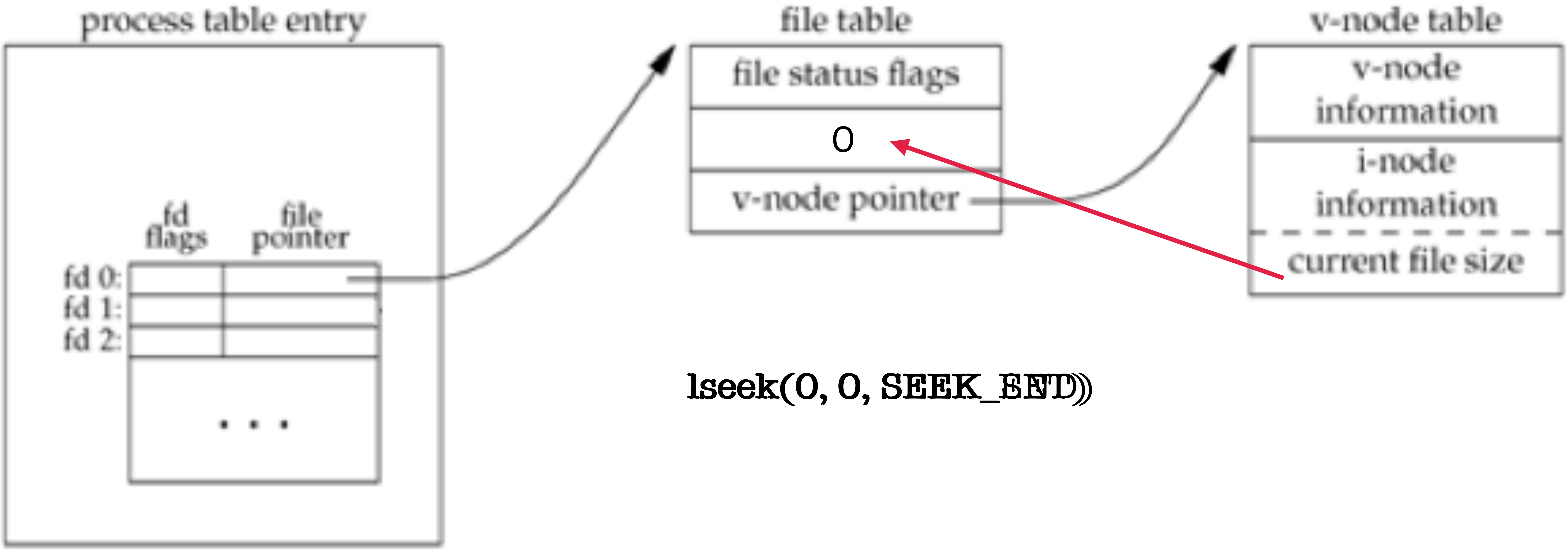
After each write(2) completes, the current file offset in the file table entry is incremented.

If the current file offset is larger than the current file size, we change the current file size in i-node table entry.



If file was opened with O_APPEND, set corresponding flag in file status flags in file table.

For each write, the current file offset is first set to the current file size from the i-node entry.



lseek(0, 0, SEEK_END))

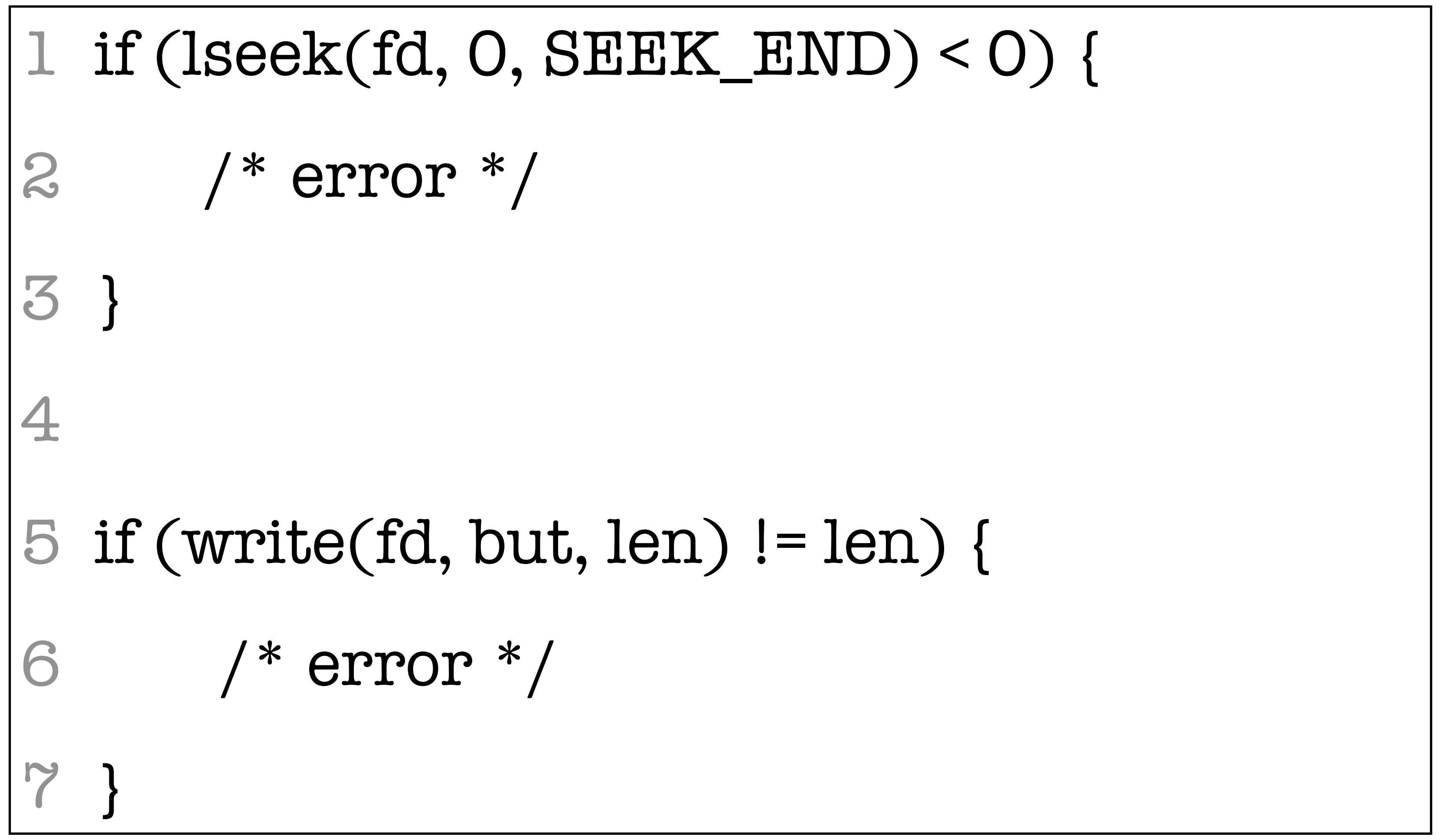
lseek(2) merely adjusts the current file offset in file table entry.

To seek to the end of a file, just copy current file size into current file offset.

To seek to the beginning of the file, simply set the offset to 0.

Atomic Operations

Early versions of Unix didn't support `O_APPEND`. Instead, you had to:



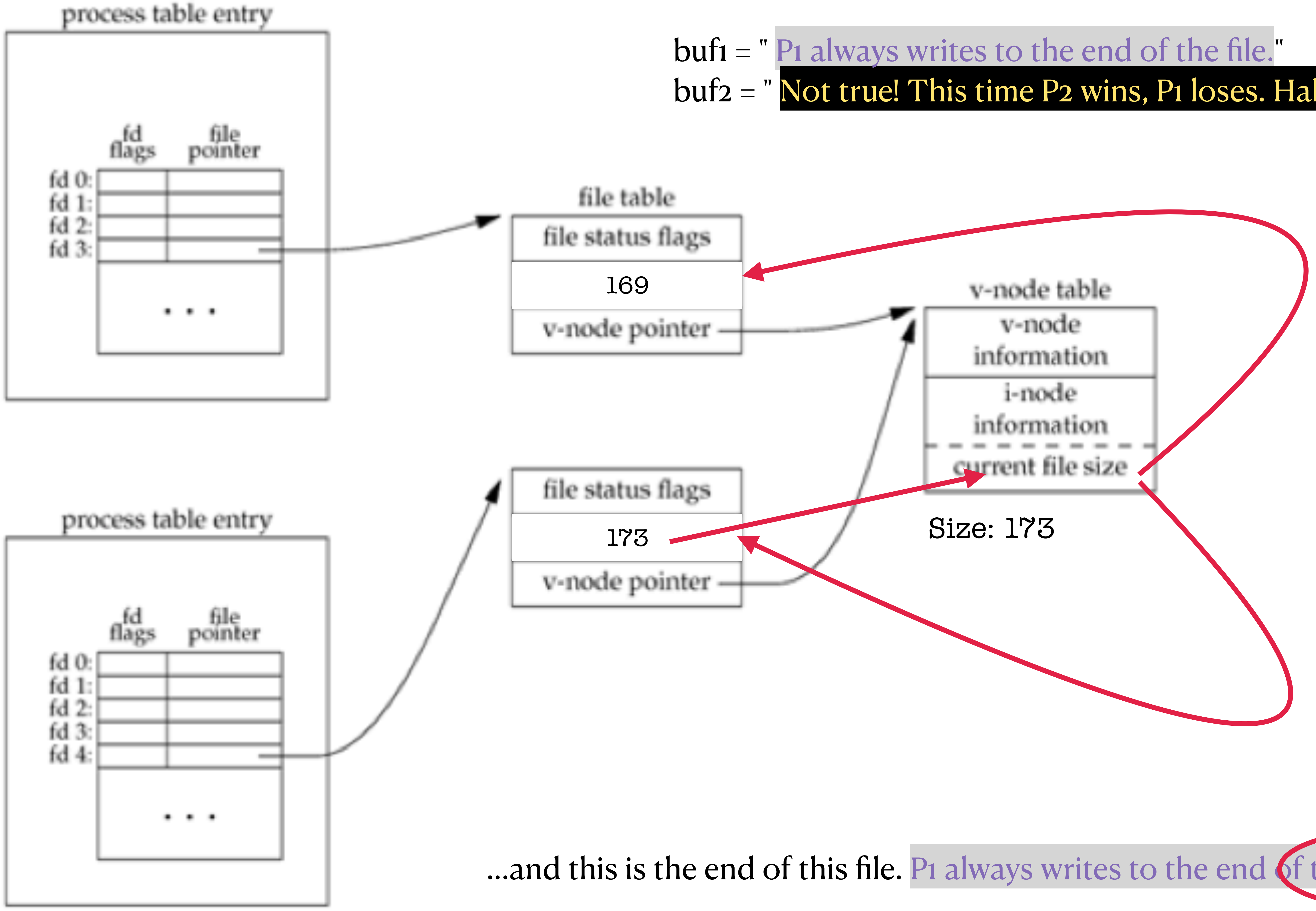
```
if (lseek(fd, 0, SEEK_END) < 0) {
        /* error */
}

if (write(fd, but, len) != len) {
        /* error */
}
```

Atomic Operations

What if another process did the same thing?

buf1 = " P1 always writes to the end of the file."

buf2 = " Not true! This time P2 wins, P1 loses. Haha!"



```
P1:
lseek(fd, 0, SEEK_END)

current offset = 128

write(fd, buf1, 41)

new offset = 128 + 41 = 169
```

```
P2:
lseek(fd, 0, SEEK_END)

current offset = 128

write(fd, buf2, 45)

new offset = 128 + 45 = 173

size = 173
```

...and this is the end of this file. P1 always writes to the end of the file.aha!

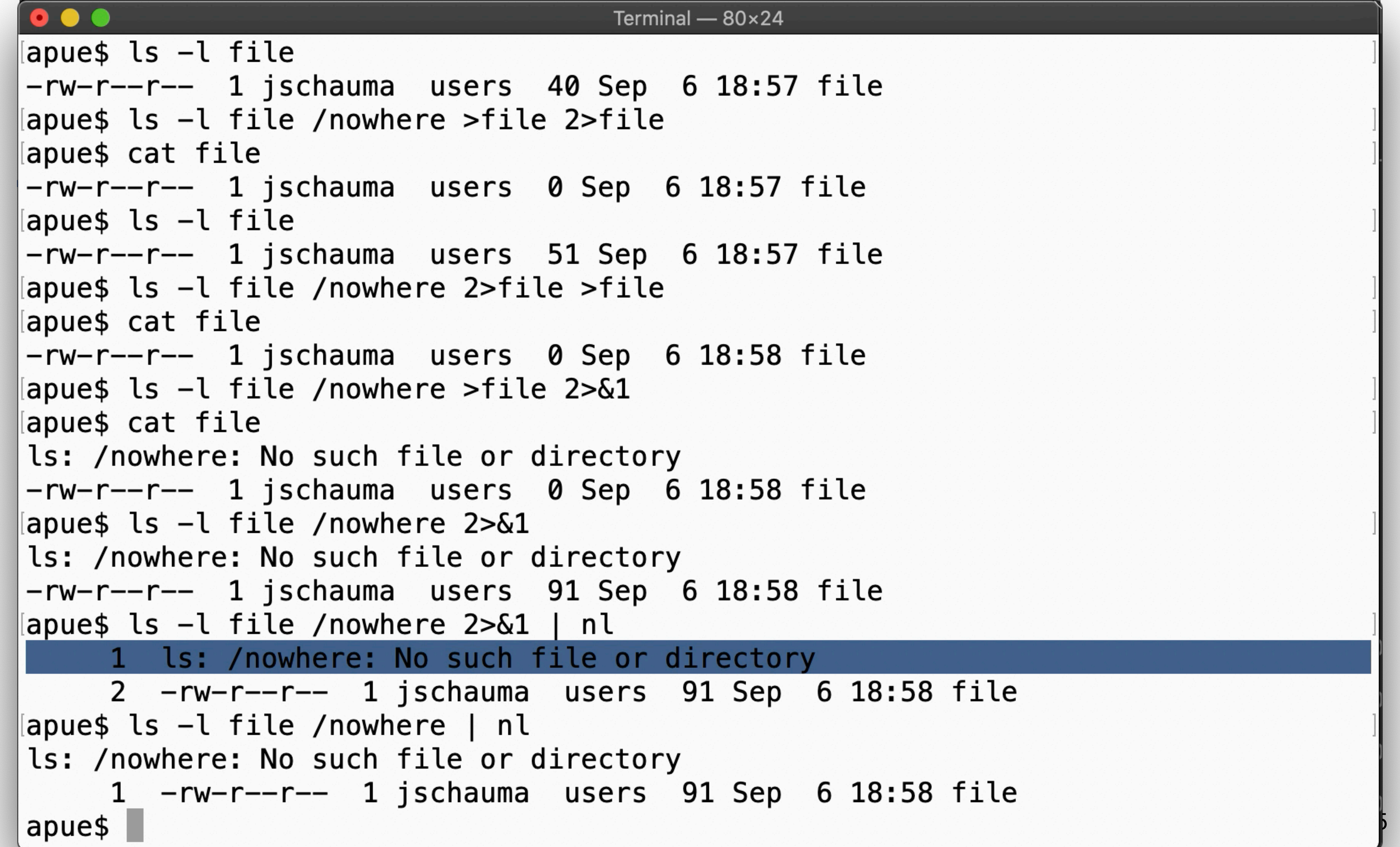
Atomic Operations

O_APPEND solves the case for writing to the end, but what if we want to write atomically anywhere else in the file?

```
#include <unistd.h>

ssize_t pread(int fd, void *buf, size_t num, off_t offset);

ssize_t write(int fd, void *buf, size_t num, off_t offset);
```

Returns: number of bytes read/written, -1 on error

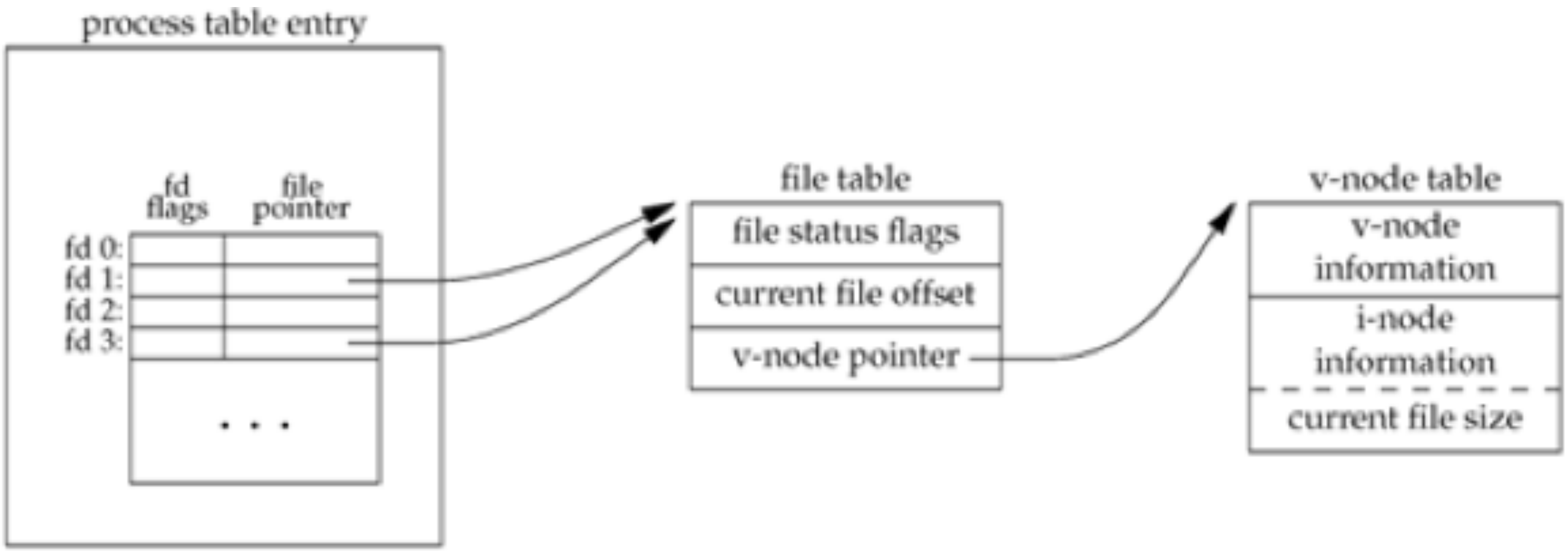
Note: current offset is not changed.

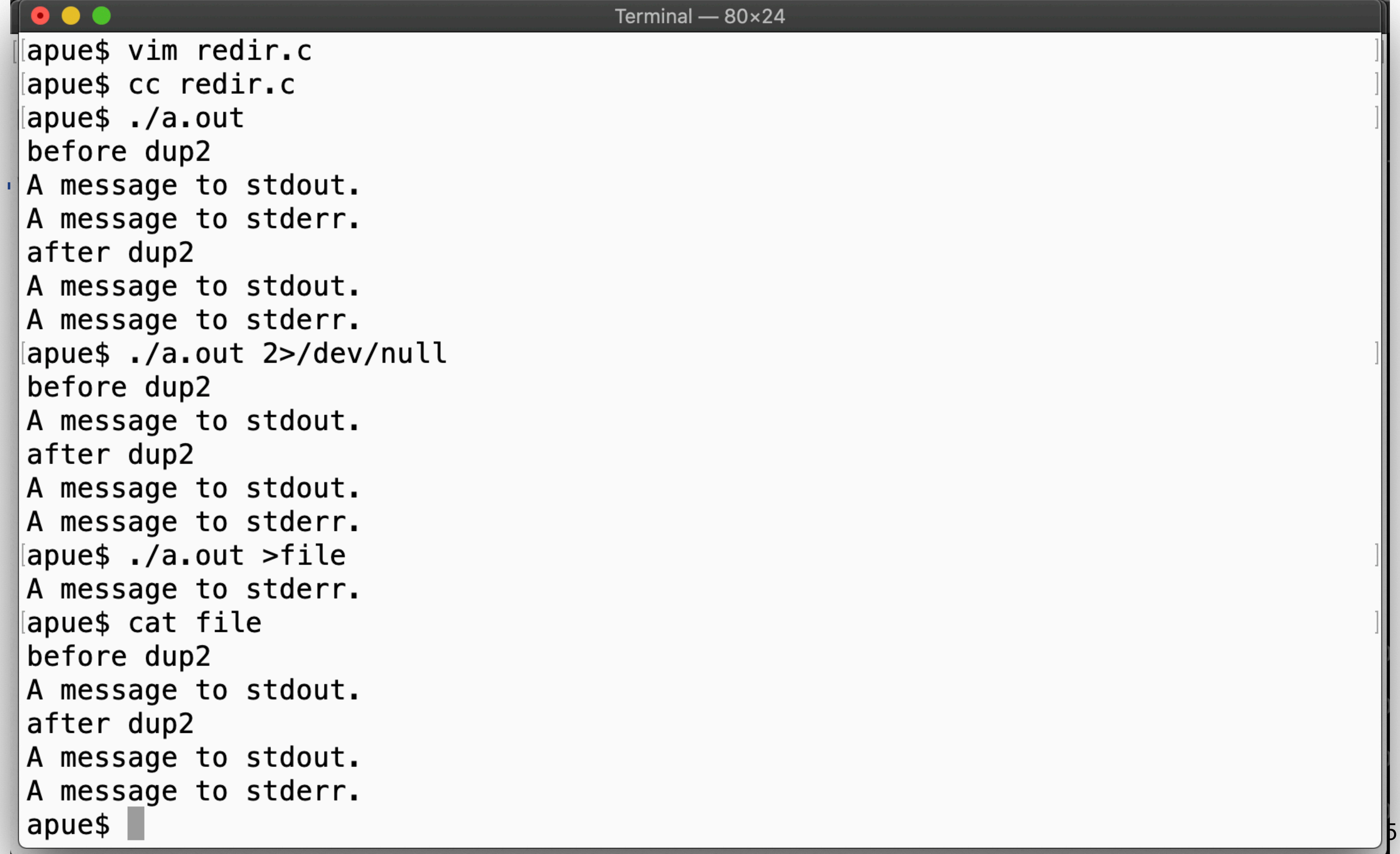


```
apue$ ls -l file
-rw-r--r--  1 jschauma  users  40 Sep  6 18:57 file
apue$ ls -l file /nowhere >file 2>file
apue$ cat file
-rw-r--r--  1 jschauma  users  0 Sep  6 18:57 file
apue$ ls -l file
-rw-r--r--  1 jschauma  users  51 Sep  6 18:57 file
apue$ ls -l file /nowhere 2>file >file
apue$ cat file
-rw-r--r--  1 jschauma  users  0 Sep  6 18:58 file
apue$ ls -l file /nowhere >file 2>&1
apue$ cat file
ls: /nowhere: No such file or directory
-rw-r--r--  1 jschauma  users  0 Sep  6 18:58 file
apue$ ls -l file /nowhere 2>&1
ls: /nowhere: No such file or directory
-rw-r--r--  1 jschauma  users  91 Sep  6 18:58 file
apue$ ls -l file /nowhere 2>&1 | nl
     1  ls: /nowhere: No such file or directory
     2  -rw-r--r--  1 jschauma  users  91 Sep  6 18:58 file
apue$ ls -l file /nowhere | nl
ls: /nowhere: No such file or directory
     1  -rw-r--r--  1 jschauma  users  91 Sep  6 18:58 file
apue$ 
```

File Descriptor Duplication

```
#include <unistd.h>

int dup(int oldfd);

int dup2(int oldfd, int newfd);

                                        Returns: newfd, -1 on error
```





```
apue$ vim redir.c
apue$ cc redir.c
apue$ ./a.out
before dup2
A message to stdout.
A message to stderr.
after dup2
A message to stdout.
A message to stderr.
apue$ ./a.out 2>/dev/null
before dup2
A message to stdout.
after dup2
A message to stdout.
A message to stderr.
apue$ ./a.out >file
A message to stderr.
apue$ cat file
before dup2
A message to stdout.
after dup2
A message to stdout.
A message to stderr.
apue$ 
```

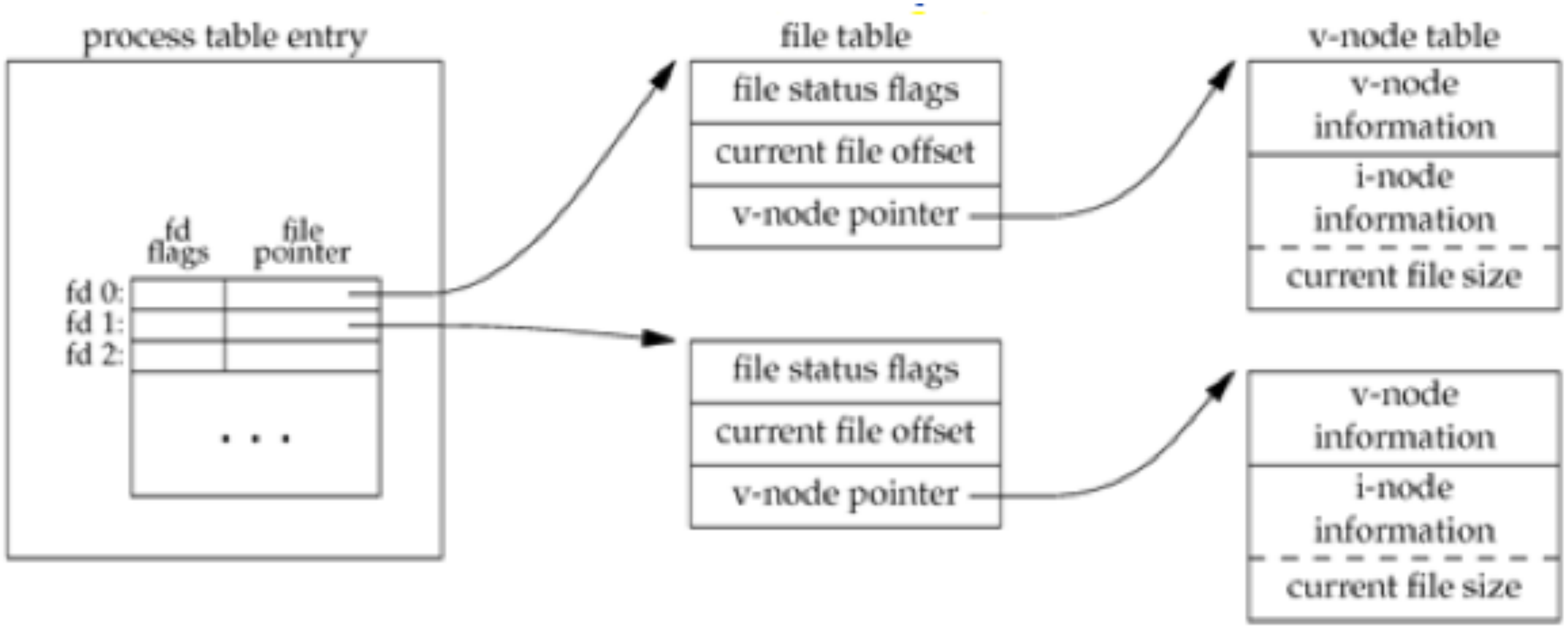
File Descriptor Control

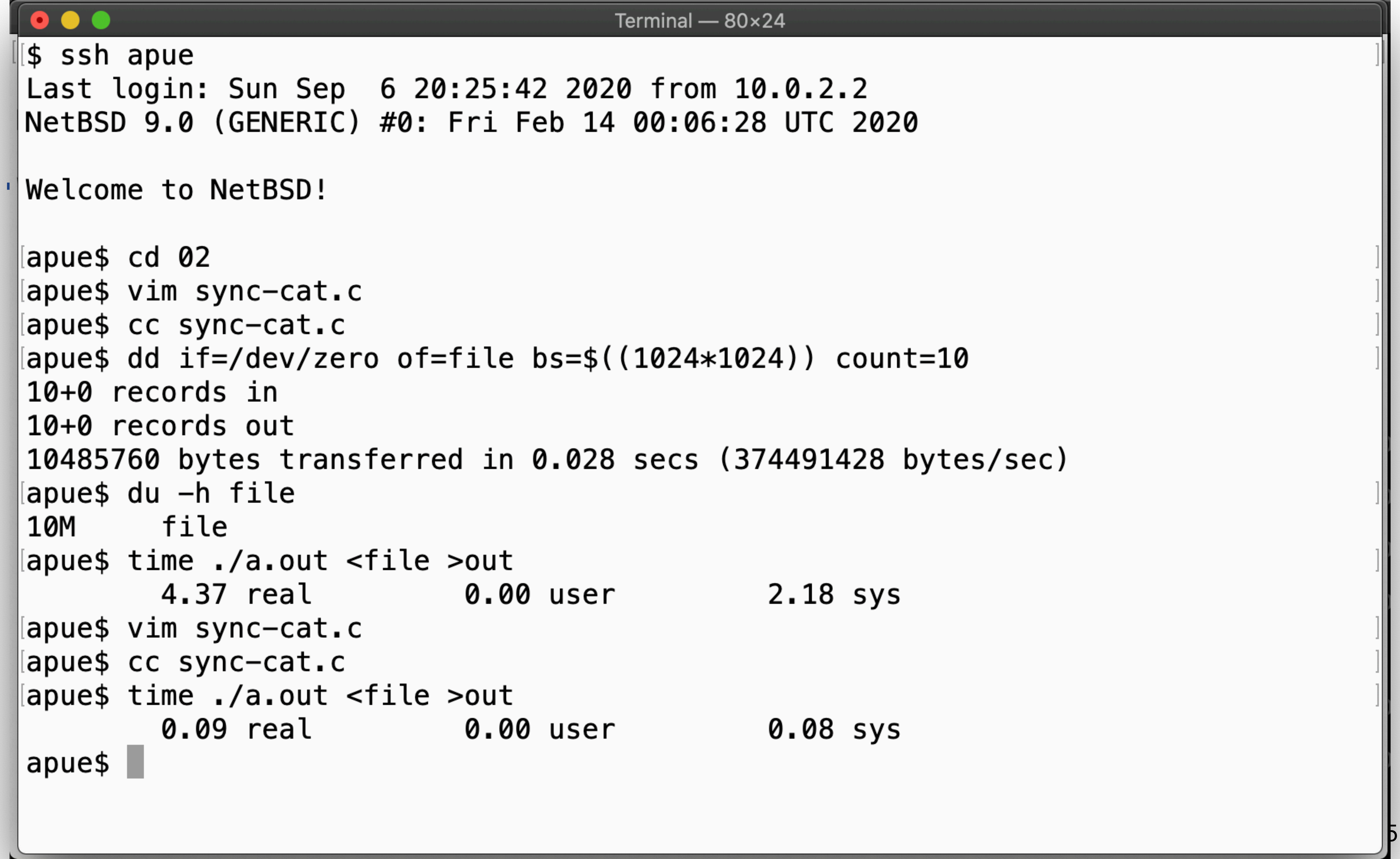
```
#include <fcntl.h>

int fcntl(int fd, int cmd, ...);
```

Returns: depends on *cmd*, -1 on error

fcntl(2) is on of those ”catch-all” functions with a myriad of purposes. Here, they all relate to changing properties of an already open file. Some of them are:

- F_DUPFD - duplicate file descriptors
- F_GETFD - get file descriptor flags
- F_SETFD - set file descriptor flags
- F_GETFL - get file status flags
- F_SETFL - set file status flags
- ...





```
$ ssh apue
Last login: Sun Sep  6 20:25:42 2020 from 10.0.2.2
NetBSD 9.0 (GENERIC) #0: Fri Feb 14 00:06:28 UTC 2020

Welcome to NetBSD!

apue$ cd 02
apue$ vim sync-cat.c
apue$ cc sync-cat.c
apue$ dd if=/dev/zero of=file bs=$((1024*1024)) count=10
10+0 records in
10+0 records out
10485760 bytes transferred in 0.028 secs (374491428 bytes/sec)
apue$ du -h file
10M     file
apue$ time ./a.out <file >out
        4.37 real         0.00 user         2.18 sys
apue$ vim sync-cat.c
apue$ cc sync-cat.c
apue$ time ./a.out <file >out
        0.09 real         0.00 user         0.08 sys
apue$ 
```

File Descriptor Control

```
#include <sys/ioctl.h>

int ioctl(int fd, unsigned long request, ...);
```

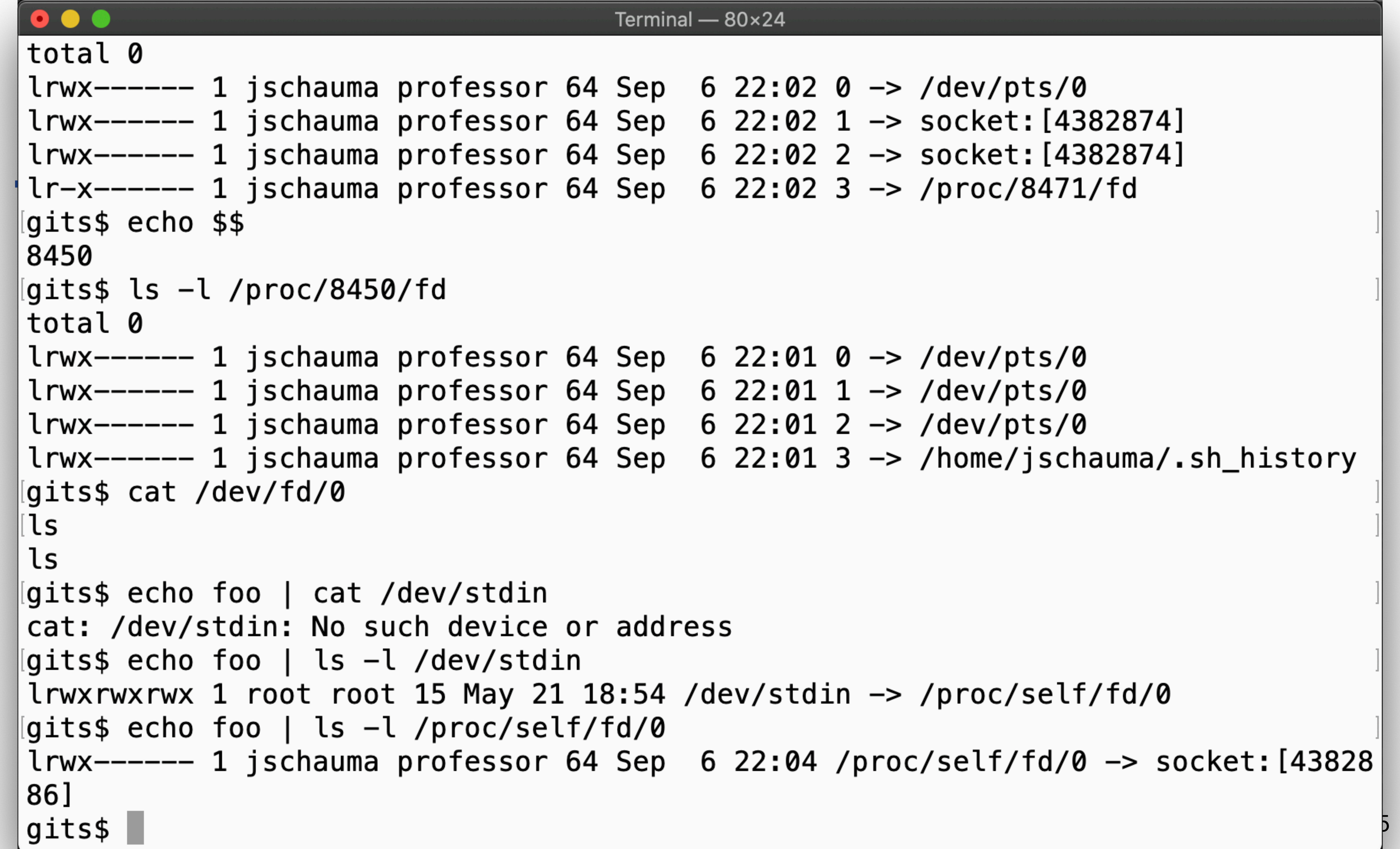
Returns: depends on *request*, -1 on error

Another catch-all function, this one is designed to handle device specifics that can't be specified via any of the previous function calls.

Examples include terminal I/O, magtape access, socket I/O, etc.

Mentioned here mostly for completeness's sake.

See also: `tty(4)`, `ioctl_list(2)` (Linux)



```
total 0
lrwx------ 1 jschauma professor 64 Sep  6 22:02 0 -> /dev/pts/0
lrwx------ 1 jschauma professor 64 Sep  6 22:02 1 -> socket:[4382874]
lrwx------ 1 jschauma professor 64 Sep  6 22:02 2 -> socket:[4382874]
lr-x------ 1 jschauma professor 64 Sep  6 22:02 3 -> /proc/8471/fd
gits$ echo $$
8450
gits$ ls -l /proc/8450/fd
total 0
lrwx------ 1 jschauma professor 64 Sep  6 22:01 0 -> /dev/pts/0
lrwx------ 1 jschauma professor 64 Sep  6 22:01 1 -> /dev/pts/0
lrwx------ 1 jschauma professor 64 Sep  6 22:01 2 -> /dev/pts/0
lrwx------ 1 jschauma professor 64 Sep  6 22:01 3 -> /home/jschauma/.sh_history
gits$ cat /dev/fd/0
ls
ls
gits$ echo foo | cat /dev/stdin
cat: /dev/stdin: No such device or address
gits$ echo foo | ls -l /dev/stdin
lrwxrwxrwx 1 root root 15 May 21 18:54 /dev/stdin -> /proc/self/fd/0
gits$ echo foo | ls -l /proc/self/fd/0
lrwx------ 1 jschauma professor 64 Sep  6 22:04 /proc/self/fd/0 -> socket:[43828
86]
gits$ 
```

Conclusion

We've covered the five syscalls on which all basic Unix I/O is based:

`open(2)`, `close(2)`, `read(2)`, `write(2)`, and `lseek(2)`

We've looked at the kernel structures used to implement these calls and discussed the impact of multiple, simultaneous processes accessing the same files.

We've seen a bunch of odd (or at least surprising) things and have written some code to clarify our understanding.

Let's put all this knowledge into practice:

`https://stevens.netmeister.org/631/f20-hw1.html`

Additional Reading

- https://en.wikipedia.org/wiki/File_descriptor
- https://en.wikipedia.org/wiki/Sparse_file
- https://tldp.org/LDP/abs/html/io-redirection.html
- https://unix.stackexchange.com/questions/98958/linux-nuisance-dev-stdin-doesnt-work-with-sockets
- https://marc.info/?l=ast-users&m=120978595414990&w=2