

### Operating system - What Does it Do.

- Operating system sits between the user and the hardware
- Operating system effectively provides a virtual machine to each user
- This virtual machine is much simpler and more convenient than real machine
- The virtual machine interface can be consistent across different hardware.
  - programs can portably access hardware across different hardware configurations
  - e.g. linux available for almost all suitable hardware
- can coordinate/share access to resources between users
- can provide privileges/security

### Operating System - What Does it Need from Hardware.

- needs hardware to provide a **privileged** mode
  - code running in privileged mode can access all hardware and memory
  - code running in privileged mode has unlimited access to memory
- needs hardware to provide a **non-privileged** mode which:
  - code running in non-privileged mode can not access hardware directly
  - code running in non-privileged mode has limited access to memory
  - provides mechanism to make requests to operating system
- operating system (kernel) code runs in **privileged** mode
- operating system runs user code in **non-privileged** mode
  - with memory access restrictions so user code can only memory allocated to it
- user code can make requests to operating system called **system calls**
  - a system call transfers execution to operating system code in privileged mode
  - at completion of request operating system (usually) returns execution back to user code in non-privileged mode

- system call allow programs to request hardware operations
- system call transfers execution to OS code in **privileged** mode
  - includes arguments specifying details of request being made
  - OS checks operation is valid & permitted
  - OS carries out operation
  - transfers execution back to user code in **non-privileged** mode
- different operating system have different system calls
  - e.g Linux system calls very different Windows system calls
- Linux provides 400+ system calls
- examples of operations that might be provided by system call:
  - read or write bytes to a file
  - request more memory
  - create a process (run a program)
  - terminate a process
  - send information via a network

## System Calls in SPIM

- SPIM provides a virtual machine which can execute MIPS programs
- SPIM also provides a tiny operating system
- small number of SPIM system calls for I/O and memory allocation
- access is via the **syscall** instruction
  - MIPS programs running on real hardware also use **syscall**
  - on Linux **syscall**, passes execution to operating system code
  - Linux operating system code carries out request specified in  $\$v0$  and  $\$a0$
- SPIM system calls are designed for students writing tiny MIPS programs without library functions
  - e.g SPIM system call **1** - print an integer, system call **5** read an integer
- system calls on real operating systems are more general
  - e.g. system call might be read n bytes, write n bytes
  - users don't normally access system calls directly
  - users call library functions e.g. **printf** & **fgets**
  - which make system calls (often via other functions)

## Experimenting with Linux System Calls

- Linux provides 400+ system calls
  - see `/usr/include/asm/unistd_64.h` on a CSE machine
- like SPIM every Linux system call has a number, e.g write bytes to a file is system call **2**
- the C function **syscall** allows to make a Linux system call without writing assembler
  - **syscall** will be written partly/entirely in assembler
    - e.g. source for x86\_64: [https://code.woboq.org/userspace/glibc/sysdeps/unix/sysv/linux/x86\\_64/syscall.S.html](https://code.woboq.org/userspace/glibc/sysdeps/unix/sysv/linux/x86_64/syscall.S.html)
- **syscall** is not normally used by programmers in regular C code
  - most system calls have their own C wrapper function
  - e.g. the write system call has a wrapper C function called **write**
  - these wrapper function are safer & more convenient
- we only use **syscall** to experiment & learn

```
// hello world implemented with a direct syscall
// This isn't portable or readable but shows us what system calls look like.
#include <unistd.h>
int main(void) {
    char bytes[16] = "Hello, Andrew!\n";
    // argument 1 to syscall is the system call number, 1 is write
    // remaining arguments are specific to each system call
    // write system call takes 3 arguments:
    // 1) file descriptor, 1 == stdout
    // 2) memory address of first byte to write
    // 3) number of bytes to write
    syscall(1, 1, bytes, 15); // prints Hello, Andrew! on stdout
    return 0;
}
```

source code for hello\_syscalls.c

## Using read & write system calls to copy stdin to stdout

```
// copy stdin to stdout with read & write syscalls
while (1) {
    char bytes[4096];
    // system call number 0 is read
    // read system call takes 3 arguments:
    // 1) file descriptor, 1 == stdin
    // 2) memory address to put bytes read
    // 3) maximum number of bytes read
    // returns number of bytes actually read
    long bytes_read = syscall(0, 0, bytes, 4096);
    if (bytes_read <= 0) {
        break;
    }
    syscall(1, 1, bytes, bytes_read); // prints bytes to stdout
}
}
```

source code for cat\_syscalls.c

## What Really are Files and Directories?

- **file systems** manage persistent stored data e.g. on magnetic disk or SSD
- On Unix-like systems:
  - a **file** is sequence (array) of zero or more bytes.
  - no meaning for bytes associated with file
    - file metadata doesn't record that it is e.g. ASCII, MP4, JPG, ...
    - Unix-like files are just bytes
  - a **directory** is an object containing zero or more files or directories.
- file systems maintain metadata for files & directories, e.g. permissions
- system calls provide operations to manipulate files.
- libc provides a low-level API to manipulate files.
- `stdio.h` provides more portable, higher-level API to manipulate files.

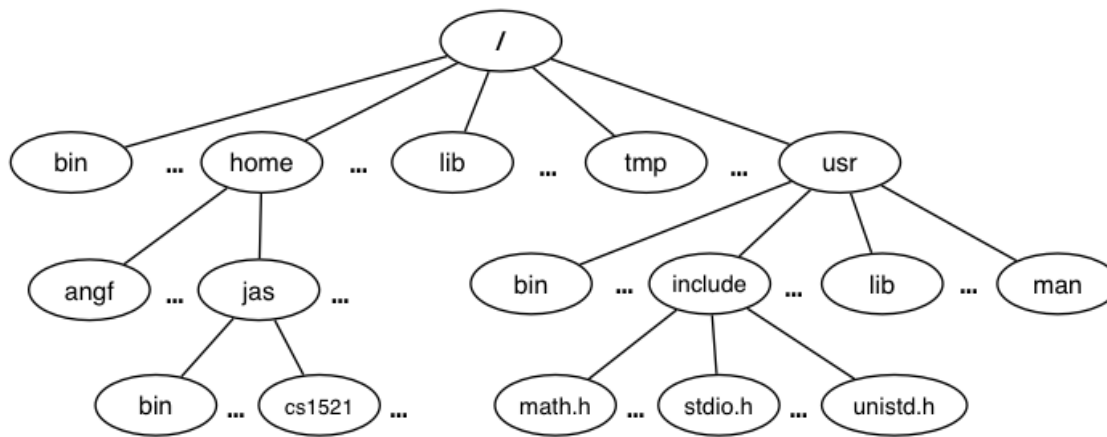
- Unix-like filenames are sequences of 1 or more bytes.
  - filenames can contain any byte except `0x00` and `0x2F`
  - `0x00` bytes (ASCII `'\0'`) used to terminate filenames
  - `0x2F` bytes (ASCII `'/'`) used to separate components of pathnames.
  - maximum filename length, depends on file system, typically 255
- Two filenames can not be used - they have a special meaning:
  - `.` current directory
  - `..` parent directory
- Some programs (shell, ls) treat filenames starting with `.` specially.
- Unix-like directories are sets of files or directories

## Unix/Linux Pathnames

- Files & directories accessed via pathnames, e.g: `/home/z5555555/lab07/main.c`
- *absolute* pathnames start with a leading `/` and give full path from root
  - e.g. `/usr/include/stdio.h`, `/cs1521/public_html/`
- every process (running program) has a *current working directory* (CWD)
  - this is an absolute pathname
- shell command `pwd` prints *current working directory*
- *relative* pathname do not start with a leading `/`
  - e.g. `../.. /another/path/prog.c`, `./a.out`, `main.c`
- *relative* pathnames appended to *current working directory* of process using them
- Assume process *current working directory* is `/home/z5555555/lab07/`
  - `main.c` translated to absolute path `/home/z5555555/lab07/main.c`
  - `../a.out` translated to absolute path `/home/z5555555/lab07/.. /a.out`
  - which is equivalent to absolute path `/home/z5555555/a.out`

## Everything is a File

- Originally files only managed data stored on a magnetic disk.
- Unix philosophy is: **Everything is a File.**
- File system used to access:
  - files
  - directories (folders)
  - storage devices (disks, SSD, ...)
  - peripherals (keyboard, mouse, USB, ...)
  - system information
  - inter-process communication
  - network
  - ...



- Unix/Linux file system is tree-like
- Exception: if you follow symbolic links it is a *graph*.
  - and you may infinitely loop attempting to traverse a file system
  - but only if you follow symbolic links

## File Metadata

Metadata for file system objects is stored in **inodes**, which hold

- location of file contents in file systems
- file type (regular file, directory, ...)
- file size in byte
- file ownership
- file access permissions - who can read, write, execute the file
- timestamps - time of creation/access/update

Note: file systems add much complexity to improve performance

- e.g. very small files might be stored in an inode itself

## File Inodes

- unix-like file systems effectively have an array of inodes
- every inode has a **inode-number** (or **i-number**)- its index in this array
- directories are effectively a list of (name, inode-number) pairs
- inode-number uniquely identify files within a filesystem
  - just a zid uniquely identifies a student within UNSW
- **ls -i** prints *inode-number*, e.g.:

```

$ ls -i file.c
109988273 file.c
$
  
```

Access to files by name proceeds (roughly) as...

- open directory and scan for *name*
- if not found, “No such file or directory”
- if found as (*name*, *i number*), access inode table `inodes[i number]`
- collect file metadata and...
  - check file access permissions given current user/group
    - if don't have required access, “Permission denied”
  - collect information about file's location and size
  - update access timestamp
- use data in inode to access file contents

## Hard Links & Symbolic Links

File system *links* allow multiple paths to access the same file

- Hard links
  - multiple names referencing the same file (inode)
  - the two entries must be on the same filesystem
  - all hard links to a file have equal status
  - file destroyed when last hard link removed
  - can not create a (extra) hard link to directories
- Symbolic links (symlinks)
  - point to another path name
  - accessing the symlink (by default) accesses the file being pointed to
  - symbolic link can point to a directory
  - symbolic link can point to a pathname on another filesystems
  - symbolic links don't have permissions (just a pointer)

## Hard Links & Symbolic Links

```
$ echo 'Hello Andrew' >hello
$ ln hello hola # create hard link
$ ln -s hello selamat # create symbolic link
$ ls -l hello hola selamat
-rw-r--r-- 2 andrewt 13 Oct 23 16:18 hello
-rw-r--r-- 2 andrewt 13 Oct 23 16:18 hola
lrwxrwxrwx 1 andrewt 5 Oct 23 16:20 selamat -> hello
$ cat hello
Hello Andrew
$ cat hola
Hello Andrew
$ cat selamat
Hello Andrew
```

Unix presents a uniform interface to file system objects

- system calls manipulate objects as a *stream of bytes*
- accessed via a **file descriptor**
  - **file descriptors** are small integers
  - index to a per-process operating system table (array)

Some important system calls:

- **open()** — open a file system object, returning a **file descriptor**
- **close()** — stop using a **file descriptor**
- **read()** — read some bytes from a file **descriptor**
- **write()** — write some bytes to a file **descriptor**
- **lseek()** — move file **descriptor** to a specified offset within a file
- **stat()** — get file system metadata for a pathname

## Using system call directly to create a file

```
// cp <file1> <file2> with syscalls, no error handling!
// system call number 2 is open, takes 3 arguments:
// 1) address of zero-terminated string containing file pathname
// 2) bitmap indicating whether to write, read, ... file
//    0x41 == write to file creating if necessary
// 3) permissions if file will be newly created
//    0644 == readable to everyone, writable by owner
long read_file_descriptor = syscall(2, argv[1], 0, 0);
long write_file_descriptor = syscall(2, argv[2], 0x41, 0644);
while (1) {
    char bytes[4096];
    long bytes_read = syscall(0, read_file_descriptor, bytes, 4096);
    if (bytes_read <= 0) {
        break;
    }
    syscall(1, write_file_descriptor, bytes, bytes_read);
}
```

source code for cp\_syscalls.c

## C Library Wrappers for System Calls

- On Unix-like systems there are C library functions corresponding to each system call,
  - e.g. open, read, write, close
  - the **syscall** function is not used in normal coding
- These functions are not portable - absent from many platforms/implementations
- POSIX standardizes some of these functions
  - some non-Unix systems provide implementations of these functions
- better to use functions from standard C library, available everywhere
  - e.g. fopen, fgets, fputc from **stdio.h**
  - on unix-like systems these will call open, read, write,
- but sometimes need to use lower level functions

Unix-like (POSIX) systems add some extra file-system-related C types in these include files:

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

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

- **off\_t** – offsets within files
  - typically **int64\_t** - signed to allow backward references
- **size\_t** – number of bytes in some object
  - typically **uint64\_t** - unsigned since objects can't have negative size
- **ssize\_t** – sizes of read/written bytes
  - typically **uint64\_t** - similar to **size\_t**, but signed to allow for error values
- **struct stat** – file system object metadata
  - stores information *about* file, not its contents
  - requires other types: `ino_t`, `dev_t`, `time_t`, `uid_t`, ...

## C library wrapper for open system call

```
int open(char *pathname, int flags)
```

- open file at **pathname**, according to **flags**
- **flags** is a bit-mask defined in `<fcntl.h>`
  - `O_RDONLY` – open for reading
  - `O_WRONLY` – open for writing
  - `O_APPEND` – append on each write
  - `O_RDWR` – open object for reading and writing
  - `O_CREAT` – create file if doesn't exist
  - `O_TRUNC` – truncate to size 0
- flags can be combined e.g. `(O_WRONLY | O_CREAT)`
- if successful, return file descriptor (small non-negative `int`)
- if unsuccessful, return `-1` and set **errno**

## C library wrapper for close system call

```
int close(int fd)
```

- release open file descriptor **fd**
- if successful, return 0
- if unsuccessful, return `-1` and set **errno**
  - could be unsuccessful if **fd** is not an open file descriptor
  - e.g. if **fd** has already been closed

An aside: removing a file e.g. via `rm`

- removes the file's entry from a directory
- but the inode and data persist until
  - all references to the inode from other directories are removed
  - all processes accessing the file `close()` their file descriptor
- after this, the inode and the space used for file contents is recycled



```
ssize_t read(int fd, void *buf, size_t count)
```

- read (up to) **count** bytes from **fd** into **buf**
  - **buf** should point to array of at least **count** bytes
  - read does (can) not check **buf** points to enough space
- if successful, number of bytes actually read is returned
- 0 returned, if no more bytes to read
- -1 returned if error and `errno` set to reason
- associated with a file descriptor is a **current position** in file
- next call to **read** will return next bytes from file
- repeated calls to reads will yield entire contents of file
- can also modify this current position with `lseek`

## C library wrapper for write system call

```
ssize_t write(int fd, const void *buf, size_t count)
```

- attempt to write **count** bytes from *buf* into stream identified by file descriptor **fd**
- if successful, number of bytes actually written is returned
- if unsuccessful, return -1 and set **errno**
- does (can) not check **buf** points to **count** bytes of data
- associated with a file descriptor is a **current position** in file
- next call to **write** will follow bytes already written
- file often created by repeated calls to write
- can also modify this current position with `lseek`

## Hello write!

```
// hello world implemented with libc
#include <unistd.h>
int main(void) {
    char bytes[16] = "Hello, Andrew!\n";
    // write takes 3 arguments:
    // 1) file descriptor, 1 == stdout
    // 2) memory address of first byte to write
    // 3) number of bytes to write
    write(1, bytes, 15); // prints Hello, Andrew! on stdout
    return 0;
}
```

source code for hello\_libc.c

```
while (1) {
    char bytes[4096];
    // system call number 0 == read
    // read system call takes 3 arguments:
    // 1) file descriptor, 1 == stdin
    // 2) memory address to put bytes read
    // 3) maximum number of bytes read
    // returns number of bytes actually read
    ssize_t bytes_read = read(0, bytes, 4096);
    if (bytes_read <= 0) {
        break;
    }
    write(1, bytes, bytes_read); // prints bytes to stdout
}
```

source code for cat\_lib.c

## Using open to copy a file

```
// open takes 3 arguments:
// 1) address of zero-terminated string containing pathname of file to open
// 2) bitmap indicating whether to write, read, ... file
// 3) permissions if file will be newly created
// 0644 == readable to everyone, writable by owner
int read_file_descriptor = open(argv[1], O_RDONLY);
int write_file_descriptor = open(argv[2], O_WRONLY | O_CREAT, 0644);
while (1) {
    char bytes[4096];
    ssize_t bytes_read = read(read_file_descriptor, bytes, 4096);
    if (bytes_read <= 0) {
        break;
    }
    write(write_file_descriptor, bytes, bytes_read);
}
```

source code for cp\_lib.c

## C library wrapper for lseek system call

`off_t lseek(int fd, off_t offset, int whence)`

- change the 'current position' in the file of **fd**
- **offset** is in units of bytes, and can be negative
- **whence** can be one of ...
  - SEEK\_SET – set file position to *Offset* from start of file
  - SEEK\_CUR – set file position to *Offset* from current position
  - SEEK\_END – set file position to *Offset* from end of file
- seeking beyond end of file leaves a gap which reads as 0's
- seeking back beyond start of file sets position to start of file

Example: `lseek(fd, 0, SEEK_END);` (move to end of file)

```
int read_file_descriptor = open(argv[1], O_RDONLY);
char bytes[1];
// move to a position 1 byte from end of file
// then read 1 byte
lseek(read_file_descriptor, -1, SEEK_END);
read(read_file_descriptor, bytes, 1);
printf("last byte of the file is 0x%02x\n", bytes[0]);
// move to a position 0 bytes from start of file
// then read 1 byte
lseek(read_file_descriptor, 0, SEEK_SET);
read(read_file_descriptor, bytes, 1);
printf("first byte of the file is 0x%02x\n", bytes[0]);
```

source code for lseek.c

## Using lseek to read bytes in the middle of a file

```
printf("first byte of the file is 0x%02x\n", bytes[0]);
// move to a position 41 bytes from start of file
// then read 1 byte
lseek(read_file_descriptor, 41, SEEK_SET);
read(read_file_descriptor, bytes, 1);
printf("42nd byte of the file is 0x%02x\n", bytes[0]);
// move to a position 58 bytes from current position
// then read 1 byte
lseek(read_file_descriptor, 58, SEEK_CUR);
read(read_file_descriptor, bytes, 1);
printf("100th byte of the file is 0x%02x\n", bytes[0]);
```

source code for lseek.c

## stdio.h - C Standard Library I/O Functions

- `stdio.h` is part of standard C library
- available in every C implementation that can do I/O
- `stdio.h` functions are portable, convenient & efficient
- use `stdio.h` functions for file operations unless you have a good reason not to
  - e.g. program with special I/O requirements like a database implementation
- on Unix-like systems they will call `open/read/write/...`
  - but with buffering for efficiency

```
FILE *fopen(const char *pathname, const char *mode)
```

- `stdio.h` equivalent to `open`
- **mode** is string of 1 or more characters including:
  - **r** open text file for reading.
  - **w** open text file for writing truncated to 0 zero length if it exists created if does not exist
  - **a** open text file for writing writes append to it if it exists created if does not exist
- `fopen` returns a **FILE\*** pointer
- **FILE** is an opaque struct - we can not access fields

```
int fclose(FILE *stream)
```

- `stdio.h` equivalent to `close`

## stdio.h - read and writing

```
int fgetc(FILE *stream)           // read a byte
int fputc(int c, FILE *stream)    // write a byte

char *fputs(char *s, FILE *stream) // write a string

char *fgets(char *s, int size, FILE *stream) // read a line

// formatted input
int fscanf(FILE *stream, const char *format, ...)
// formatted output
int fprintf(FILE *stream, const char *format, ...)

// read array of bytes (fgetc + loop often better)
size_t fread(void *ptr, size_t size, size_t nmemb, FILE *stream);
// write array of bytes (fputc + loop often better)
size_t fwrite(const void *ptr, size_t size, size_t nmemb,
              FILE *stream)
```

## stdio.h - using fputc to output bytes

```
char bytes[] = "Hello, stdio!\n"; // 15 bytes
// write 14 bytes so we don't write (terminating) 0 byte
for (int i = 0; i < (sizeof bytes) - 1; i++) {
    fputc(bytes[i], stdout);
}
// or as we know bytes is 0-terminated
for (int i = 0; bytes[i] != '\0'; i++) {
    fputc(bytes[i], stdout);
}
// or if you prefer pointers
for (char *p = &bytes[0]; *p != '\0'; p++) {
    fputc(*p, stdout);
}
```

source code for `hello_stdio.c`

```
char bytes[] = "Hello, stdio!\n"; // 15 bytes
```

```
// fputs relies on bytes being 0-terminated  
fputs(bytes, stdout);  
// write 14 1 byte items  
fwrite(bytes, 1, (sizeof bytes) - 1, stdout);  
// %s relies on bytes being 0-terminated  
fprintf(stdout, "%s", bytes);
```

source code for hello\_stdio.c

## stdio.h - using fgetc to copy stdin to stdout

```
// c can not be char (common bug)  
// fgetc returns 0..255 and EOF (usually -1)  
int c;  
// return bytes from the stream (stdin) one at a time  
while ((c = fgetc(stdin)) != EOF) {  
    fputc(c, stdout); // write the byte to standard output  
}
```

source code for cat\_fgetc.c

## stdio.h - using fgets to copy stdin to stdout

```
// return bytes from the stream (stdin) line at a time  
char line[256];  
while (fgets(line, sizeof line, stdin) != NULL) {  
    fputs(line, stdout);  
}  
//  
// NOTE: fgets returns a null-terminated string  
//         in other words a 0 byte marks the end of the bytes read  
//  
// fgets can not be used to read bytes which are 0  
// fputs takes a null-terminated string  
// so fputs can not be used to write bytes which are 0  
// hence you can't use fget/fputs for binary data e.g. jpgs
```

source code for cat\_fgets.c

```
// BUFSIZ is defined in stdio.h - its an efficient value to use
// but any value would work
while (1) {
    char bytes[BUFSIZ];
    ssize_t bytes_read = fread(bytes, 1, sizeof bytes, stdin);
    if (bytes_read <= 0) {
        break;
    }
    fwrite(bytes, 1, bytes_read, stdout);
}
```

source code for cat\_fwrite.c

## stdio.h - creating a file

```
// create file "hello.txt" containing 1 line: Hello, Andrew
#include <stdio.h>
#include <stdlib.h>
int main(int argc, char *argv[]) {
    FILE *output_stream = fopen("hello.txt", "w");
    if (output_stream == NULL) {
        perror("hello.txt");
        return 1;
    }
    fprintf(output_stream, "Hello, Andrew!\n");
    // fclose will flush data to file
    // best to close file ASAP
    // but doesn't matter as file automatically closed on exit
    fclose(output_stream);
    return 0;
}
```

source code for create\_file\_fopen.c

## stdio.h - using fgetc to copy a file

```
FILE *input_stream = fopen(argv[1], "rb");
if (input_stream == NULL) {
    perror(argv[1]); // prints why the open failed
    return 1;
}
FILE *output_stream = fopen(argv[2], "wb");
if (output_stream == NULL) {
    perror(argv[2]);
    return 1;
}
int c; // not char!
while ((c = fgetc(input_stream)) != EOF) {
    fputc(c, output_stream);
}
fclose(input_stream); // optional as close occurs
fclose(output_stream); // automatically on exit
```

source code for cp\_fgetc.c

```
FILE *input_stream = fopen(argv[1], "rb");
```

```
FILE *output_stream = fopen(argv[2], "wb");
```

```
// this will be slightly faster than an a fgetc/fputc loop
while (1) {
    char bytes[BUFSIZ];
    size_t bytes_read = fread(bytes, 1, sizeof bytes, input_stream);
    if (bytes_read <= 0) {
        break;
    }
    fwrite(bytes, 1, bytes_read, output_stream);
}
fclose(input_stream); // optional as close occurs
fclose(output_stream); // automatically on exit
```

source code for cp\_fwrite.c

## stdio.h - other operations

```
int fseek(FILE *stream, long offset, int whence);
```

- **fseek** is stdio equivalent to `lseek`
- like `lseek` **offset** can be positive or negative
- like `lseek` **whence** can be `SEEK_SET`, `SEEK_CUR` or `SEEK_END` making **offset** relative to file start, current position or file end

```
int fflush(FILE *stream);
```

- flush any buffered data on output stream

## Using fseek to read the last byte then the first byte of a file

```
FILE *input_stream = fopen(argv[1], "rb");
// move to a position 1 byte from end of file
// then read 1 byte
fseek(input_stream, -1, SEEK_END);
printf("last byte of the file is 0x%02x\n", fgetc(input_stream));
// move to a position 0 bytes from start of file
// then read 1 byte
fseek(input_stream, 0, SEEK_SET);
printf("first byte of the file is 0x%02x\n", fgetc(input_stream));
```

source code for fseek.c

- NOTE: important error checking is missing above

```
// move to a position 41 bytes from start of file
// then read 1 byte
fseek(input_stream, 41, SEEK_SET);
printf("42nd byte of the file is 0x%02x\n", fgetc(input_stream));
// move to a position 58 bytes from current position
// then read 1 byte
fseek(input_stream, 58, SEEK_CUR);
printf("100th byte of the file is 0x%02x\n", fgetc(input_stream));
```

source code for fseek.c

- NOTE: important error checking is missing above

## Using fseek to change a random file bit

```
FILE *f = fopen(argv[1], "r+"); // open for reading and writing
fseek(f, 0, SEEK_END); // move to end of file
long n_bytes = ftell(f); // get number of bytes in file
srandom(time(NULL)); // initialize random number
// generator with current time

long target_byte = random() % n_bytes; // pick a random byte
fseek(f, target_byte, SEEK_SET); // move to byte
int byte = fgetc(f); // read byte
int bit = random() % 8; // pick a random bit
int new_byte = byte ^ (1 << bit); // flip the bit
fseek(f, -1, SEEK_CUR); // move back to same position
fputc(new_byte, f); // write the byte
fclose(f);
```

source code for fuzz.c

- random changes to search for errors/vulnerabilities called fuzzing

## Using fseek to create a gigantic sparse file (advanced topic)

```
// Create a 16 terabyte sparse file
// https://en.wikipedia.org/wiki/Sparse_file
// error checking omitted for clarity
#include <stdio.h>
int main(void) {
    FILE *f = fopen("sparse_file.txt", "w");
    fprintf(f, "Hello, Andrew!\n");
    fseek(f, 16L * 1000 * 1000 * 1000 * 1000, SEEK_CUR);
    fprintf(f, "Goodbye, Andrew!\n");
    fclose(f);
    return 0;
}
```

source code for create\_gigantic\_file.c

- almost all the 16Tb are zeros which the file system doesn't actually store



- as we often read/write to stdin/stdout `stdio.h` provides convenience functions, we can use:

```
int getchar()      // fgetc(stdin)
int putchar(int c) // fputc(c, stdin)

int puts(char *s) // fputs(s, stdout)

int scanf(char *format, ...) // fscanf(stdin, format, ...)
int printf(char *format, ...) // fprintf(stdout, format, ...)

char *gets(char *s); // NEVER USE
```

## stdio.h - I/O to strings

`stdio.h` provides useful functions which operate on strings

```
int snprintf(char *str, size_t size, const char *format, ...);
```

- like `printf`, but output goes to char array **str**
- handy for creating strings passed to other functions
- do not use unsafe related function: `sprintf`

```
int sscanf(const char *str, const char *format, ...);
```

- like `scanf`, but input comes from char array **str**

```
int sprintf(char *str, const char *format, ...); // DO NOT USE
```

- like **`sprintf`** but dangerous because can overflow **str**

## C library wrapper for stat system call

```
int stat(const char *pathname, struct stat *statbuf)
```

- returns metadata associated with **pathname** in **statbuf**
- metadata returned includes:
  - inode number
  - type (file, directory, symbolic link, device)
  - size of file in bytes (if it is a file)
  - permissions (read, write, execute)
  - times of last access/modification/status-change
- returns `-1` and sets **errno** if metadata not accessible

```
int fstat(int fd, struct stat *statbuf)
```

- same as `stat()` but gets data via an open file descriptor

```
int lstat(const char *pathname, struct stat *statbuf)`
```

- same as `stat()` but doesn't follow symbolic links

```

struct stat {
    dev_t      st_dev;           /* ID of device containing file */
    ino_t      st_ino;          /* Inode number */
    mode_t     st_mode;         /* File type and mode */
    nlink_t    st_nlink;        /* Number of hard links */
    uid_t      st_uid;          /* User ID of owner */
    gid_t      st_gid;          /* Group ID of owner */
    dev_t      st_rdev;         /* Device ID (if special file) */
    off_t      st_size;         /* Total size, in bytes */
    blksize_t  st_blksize;      /* Block size for filesystem I/O */
    blkcnt_t   st_blocks;       /* Number of 512B blocks allocated */
    struct timespec st_atim;     /* Time of last access */
    struct timespec st_mtim;     /* Time of last modification */
    struct timespec st_ctim;     /* Time of last status change */
};

```

## st\_mode field of struct stat

**st\_mode** is a bitwise-or of these values (& others):

S_IFLNK	0120000	symbolic link
S_IFREG	0100000	regular file
S_IFBLK	0060000	block device
S_IFDIR	0040000	directory
S_IFCHR	0020000	character device
S_IFIFO	0010000	FIFO
S_IRUSR	0000400	owner has read permission
S_IWUSR	0000200	owner has write permission
S_IXUSR	0000100	owner has execute permission
S_IRGRP	0000040	group has read permission
S_IWGRP	0000020	group has write permission
S_IXGRP	0000010	group has execute permission
S_IROTH	0000004	others have read permission
S_IWOTH	0000002	others have write permission
S_IXOTH	0000001	others have execute permission

## Using stat

```

struct stat s;
if (stat(pathname, &s) != 0) {
    perror(pathname);
    exit(1);
}
printf("ino = %10ld # Inode number\n", s.st_ino);
printf("mode = %10o # File mode \n", s.st_mode);
printf("nlink =%10ld # Link count \n", (long)s.st_nlink);
printf("uid = %10u # Owner uid\n", s.st_uid);
printf("gid = %10u # Group gid\n", s.st_gid);
printf("size = %10ld # File size (bytes)\n", (long)s.st_size);
printf("mtime =%10ld # Modification time (seconds since 1/1/70)\n",
      (long)s.st_mtime);

```

source code for stat.c

```
int mkdir(const char *pathname, mode_t mode)
```

- create a new directory called **pathname** with permissions **mode**
- if **pathname** is e.g. a/b/c/d
  - all of the directories a, b and c must exist
  - directory c must be writeable to the caller
  - directory d must not already exist
- the new directory contains two initial entries
  - . is a reference to itself
  - .. is a reference to its parent directory
- returns 0 if successful, returns -1 and sets errno otherwise

for example:

```
mkdir("newDir", 0755);
```

## Example of using mkdir to create directories

```
#include <stdio.h>
#include <sys/stat.h>
// create the directories specified as command-line arguments
int main(int argc, char *argv[]) {
    for (int arg = 1; arg < argc; arg++) {
        if (mkdir(argv[arg], 0755) != 0) {
            perror(argv[arg]); // prints why the mkdir failed
            return 1;
        }
    }
    return 0;
}
```

source code for mkdir.c

## Other useful Linux (POSIX) functions

```
chmod(char *pathname, mode_t mode) // change permission of file/...
```

```
unlink(char *pathname) // remove a file/directory/...
```

```
rename(char *oldpath, char *newpath) // rename a file/directory
```

```
chdir(char *path) // change current working directory
```

```
getcwd(char *buf, size_t size) // get current working directory
```

```
link(char *oldpath, char *newpath) // create hard link to a file
```

```
symlink(char *target, char *linkpath) // create a symbolic link
```

- file permissions are separated into three types:
  - **\*\*read\*** - permission to get bytes of file
  - **\*\*write\*** - permission to change bytes of file
  - **\*\*execute\*** - permission to execute file
- read/write/execute often represented as bits of an octal digit
- file permissions are specified for 3 groups of users:
  - **owner** - permissions for the file owner
  - **group** - permissions for users in the group of the file
  - **other** - permissions for any other user

## changing file permissions

```
// first argument is mode in octal
mode_t mode = strtol(argv[1], &end, 8);
// check first argument was a valid octal number
if (argv[1][0] == '\\0' || end[0] != '\\0') {
    fprintf(stderr, "%s: invalid mode: %s\n", argv[0], argv[1]);
    return 1;
}
for (int arg = 2; arg < argc; arg++) {
    if (chmod(argv[arg], mode) != 0) {
        perror(argv[arg]); // prints why the chmod failed
        return 1;
    }
}
}
```

source code for chmod.c

## removing files

```
int main(int argc, char *argv[]) {
    for (int arg = 1; arg < argc; arg++) {
        if (unlink(argv[arg]) != 0) {
            perror(argv[arg]); // prints why the unlink failed
            return 1;
        }
    }
    return 0;
}
```

source code for rm.c

```
$ gcc rm.c
$ ./a.out rm.c
$ ls -l rm.c
ls: cannot access 'rm.c': No such file or directory
```

```

int main(int argc, char *argv[]) {
    if (argc != 3) {
        fprintf(stderr, "Usage: %s <old-filename> <new-filename>\n",
            argv[0]);
        return 1;
    }
    char *old_filename = argv[1];
    char *new_filename = argv[2];
    if (rename(old_filename, new_filename) != 0) {
        fprintf(stderr, "%s rename %s %s:", argv[0], old_filename,
            new_filename);
        perror("");
        return 1;
    }
    return 0;
}

```

source code for rename.c

## cd-ing up one directory at a time

```

// use repeated chdir("../") to climb to root of the file system
char pathname[PATH_MAX];
while (1) {
    if (getcwd(pathname, sizeof pathname) == NULL) {
        perror("getcwd");
        return 1;
    }
    printf("getcwd() returned %s\n", pathname);
    if (strcmp(pathname, "/") == 0) {
        return 0;
    }
    if (chdir("../") != 0) {
        perror("chdir");
        return 1;
    }
}

```

source code for getcwd.c

## making a 1000-deep directory

```

for (int i = 0; i < 1000; i++) {
    char dirname[256];
    snprintf(dirname, sizeof dirname, "d%d", i);
    if (mkdir(dirname, 0755) != 0) {
        perror(dirname);
        return 1;
    }
    if (chdir(dirname) != 0) {
        perror(dirname);
        return 1;
    }
    char pathname[1000000];
    if (getcwd(pathname, sizeof pathname) == NULL) {
        perror("getcwd");
        return 1;
    }
    printf("\nCurrent directory now: %s\n", pathname);
}

```

source code for nest\_directories.c

## creating 1000 hard links to a file (creating the file)

```
int main(int argc, char *argv[]) {
    char pathname[256] = "hello.txt";
    // create a target file
    FILE *f1;
    if ((f1 = fopen(pathname, "w")) == NULL) {
        perror(pathname);
        return 1;
    }
    fprintf(f1, "Hello Andrew!\n");
    fclose(f1);
}
```

source code for many\_links.c

## creating 1000 hard links to a file (checking the file)

```
for (int i = 0; i < 1000; i++) {
    printf("Verifying '%s' contains: ", pathname);
    FILE *f2;
    if ((f2 = fopen(pathname, "r")) == NULL) {
        perror(pathname);
        return 1;
    }
    int c;
    while ((c = fgetc(f2)) != EOF) {
        fputc(c, stdout);
    }
    fclose(f2);
}
```

source code for many\_links.c

## creating 1000 hard links to a file (creating a link)

```
char new_pathname[256];
snprintf(new_pathname, sizeof new_pathname,
         "hello_%d.txt", i);
printf("Creating a link %s -> %s\n",
       new_pathname, pathname);
if (link(pathname, new_pathname) != 0) {
    perror(pathname);
    return 1;
}
}
return 0;
}
```

source code for many\_links.c

```

#include <sys/types.h>
#include <dirent.h>

// open a directory stream for directory name
DIR *opendir(const char *name);

// return a pointer to next directory entry
struct dirent *readdir(DIR *dirp);

// close a directory stream
int closedir(DIR *dirp);

```

## Using opendir/readdir to print directory contents (advanced)

```

for (int arg = 1; arg < argc; arg++) {
    DIR *dirp = opendir(argv[arg]);
    if (dirp == NULL) {
        perror(argv[arg]); // prints why the open failed
        return 1;
    }
    struct dirent *de;
    while ((de = readdir(dirp)) != NULL) {
        printf("%ld %s\n", de->d_ino, de->d_name);
    }
    closedir(dirp);
}

```

source code for list\_directory.c

## writing an array as binary data (using fwrite)

```

int array[10] = { 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 };
FILE *f = fopen("array.save", "w");
if (f == NULL) {
    perror("array.save");
    return 1;
}
// assuming int are 4 bytes, this will
// write 40 bytes of array to "array.save"
if (fwrite(array, 1, sizeof array, f) != sizeof array) {
    perror("array.save");
    return 1;
}
fclose(f);

```

source code for write\_array.c

## reading an array as binary data (using fread)

```
int array[10];
FILE *f = fopen("array.save", "r");
if (f == NULL) {
    perror("array.save");
    return 1;
}
// read array: NOT-PORTABLE: depends on size of int and byte-order
if (fread(array, 1, sizeof array, f) != sizeof array) {
    perror("array.save");
    return 1;
}
fclose(f);
for (int i = 0; i < 10; i++) {
    printf("%d ", array[i]);
}
printf("\n");
```

source code for read\_array.c

## writing a pointer as binary data (using fwrite)

```
int array[10] = { 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 };
int *p = &array[5];
FILE *f = fopen("array.save", "w");

if (fwrite(array, 1, sizeof array, f) != sizeof array) {
    perror("array.save");
    return 1;
}
if (fwrite(&p, 1, sizeof p, f) != sizeof p) {
    perror("array.save");
    return 1;
}
fclose(f);
```

source code for write\_pointer.c

## reading a pointer as binary data (using fread)

```
int array[10];
int *p;
FILE *f = fopen("array.save", "r");

if (fread(array, 1, sizeof array, f) != sizeof array) {
    perror("array.save");
    return 1;
}
// BROKEN - address of array has almost certainly changed
// BROKEN - so address p needs to point has changed
if (fread(&p, 1, sizeof p, f) != sizeof p) {
    perror("array.save");
    return 1;
}
fclose(f);
```

source code for read\_pointer.c



```

int read_file_descriptor = open(argv[1], O_RDONLY);
int write_file_descriptor = open(argv[2], O_WRONLY | O_CREAT, 0644);
// copy bytes 1 at a time
while (1) {
    char bytes[1];
    ssize_t bytes_read = read(read_file_descriptor, bytes, 1);
    if (bytes_read <= 0) {
        break;
    }
    write(write_file_descriptor, bytes, 1);
}

```

source code for cp\_libc\_one\_byte.c

- similar to earlier example [source code for cp\\_libc.c](#) but one byte at time

## I/O Performance & Buffering - Copying One Byte Per Time

```

$ clang -O3 cp_libc_one_byte.c -o cp_libc_one_byte
$ dd bs=1M count=10 </dev/urandom >random_file
10485760 bytes (10 MB, 10 MiB) copied, 0.183075 s, 57.3 MB/s
$ time ./cp_libc_one_byte random_file random_file_copy
real 0m5.262s
user 0m0.432s
sys 0m4.826s

```

- much slower than previous version which copies 4096 bytes at a time

```

$ clang -O3 cp_libc.c -o cp_libc
$ time ./cp_libc random_file random_file_copy
real 0m0.008s
user 0m0.001s
sys 0m0.007s

```

- main reason - system calls are expensive

## I/O Performance & Buffering - stdio Copying 1 Byte Per Time

```

FILE *input_stream = fopen(argv[1], "rb");
if (input_stream == NULL) {
    perror(argv[1]); // prints why the open failed
    return 1;
}
FILE *output_stream = fopen(argv[2], "wb");
if (output_stream == NULL) {
    perror(argv[2]);
    return 1;
}
int c; // not char!
while ((c = fgetc(input_stream)) != EOF) {
    fputc(c, output_stream);
}
fclose(input_stream); // optional as close occurs
fclose(output_stream); // automatically on exit

```

source code for cp\_fgetc.c

```
$ clang -O3 cp_fgetc.c -o cp_fgetc
$ time ./cp_fgetc random_file random_file_copy
real 0m0.059s
user 0m0.042s
sys 0m0.009s
```

- at the user level copies 1 byte at time using `fgetc/fputc`
- much faster than copying 1 byte at time using `read/write`
- little slower than copying 4096 bytes at time using `read/write`
- how?

## I/O Performance & Buffering - stdio buffering

- assume stdio buffering size (`BUFSIZ`) is 4096 (typical)
- stdio **buffers** 1 byte `fgetc/fputc` into 4096 bytes `read/write`
- first `fgetc` reads 4096 bytes into an array (input **buffer**)
  - next 4095 `fgetc` calls get byte from array
- first 4095 `fputc` put bytes into another array (output **buffer**)
  - next 4095 `fgetc` get byte from array
- output buffer\* emptied by **exit** or `main` returning
- data in output buffer
- program can force empty of output buffer with **fflush** call

## reimplementing stdio.h no buffering - struct

```
// re-implementation of stdio functions fopen, fgetc, fputc, fclose
// with no buffering and *zero* error handling for clarity
#include <unistd.h>
#include <sys/types.h>
#include <sys/stat.h>
#include <fcntl.h>
#include <stdint.h>
#include <stdlib.h>
#include <assert.h>
#include <stdio.h>
#define MY_EOF -1
// struct to hold data for a stream
typedef struct my_file {
    int fd;
} my_file_t;
```

source code for `cp_unbuffered.c`

```

my_file_t *my_fopen(char *file, char *mode) {
    int fd = -1;
    if (mode[0] == 'r') {
        fd = open(file, O_RDONLY);
    } else if (mode[0] == 'w') {
        fd = open(file, O_WRONLY | O_CREAT, 0666);
    } else if (mode[0] == 'a') {
        fd = open(file, O_WRONLY | O_APPEND);
    }
    if (fd == -1) {
        return NULL;
    }
    my_file_t *f = malloc(sizeof *f);
    f->fd = fd;
    return f;
}

```

source code for cp\_unbuffered.c

## reimplementing stdio.h no buffering - - my\_fgetc

```

int my_fgetc(my_file_t *f) {
    uint8_t byte;
    int bytes_read = read(f->fd, &byte, 1);
    if (bytes_read == 1) {
        return byte;
    } else {
        return MY_EOF;
    }
}

```

source code for cp\_unbuffered.c

## reimplementing stdio.h no buffering - - my\_fputc

```

int my_fputc(int c, my_file_t *f) {
    uint8_t byte = c;
    if (write(f->fd, &byte, 1) == 1) {
        return byte;
    } else {
        return MY_EOF;
    }
}

```

source code for cp\_unbuffered.c

```
int my_fclose(my_file_t *f) {  
    int result = close(f->fd);  
    free(f);  
    return result;  
}
```

source code for cp\_unbuffered.c

## reimplementing stdio.h - buffering (advanced topic)

- reimplementing stdio with input buffering

source code for cp\_input\_buffered.c

- and output buffering

source code for cp\_output\_buffered.c

## File System Summary

Operating systems provide a *file system*

- as an abstraction over physical storage devices (e.g. disks)
- providing named access to chunks of related data (files)
- providing access (sequential/random) to the contents of files
- allowing files to be arranged in a hierarchy of directories
- providing control over access to files and directories
- managing other metadata associated with files (size, location, ...)

Operating systems also manage other resources

- memory, processes, processor time, i/o devices, networking, ...