CSCE 313-200
Introduction to Computer Systems
Spring 2022

File System
Dmitri Loguinov
Texas A&M University

March 22, 2022
Chapter 11: Roadmap

11.1 I/O devices
11.2 I/O function
11.3 OS design issues
11.4 I/O buffering
11.5 Disk scheduling
11.6 RAID
11.7 Disk cache
11.8-11.10 Unix, Linux, Windows
I/O Devices

• I/O usually refers to physical devices
  – Such as disk, network card, printer, keyboard

• Almost every device in the system is I/O
  – Except RAM, CPU, and possibly certain chipsets built into the motherboard

• Transfer of data between devices and RAM thru DMA
I/O Devices

• How fast is I/O compared to RAM speed?
  - Usually slow, but it depends…
• How to measure speed?
  - Kbps, Mbps, Gbps refer to bits/sec
  - KB/s, MB/s, GB/s refer to bytes/sec
• Use a notation with K = 1000 bits/bytes
I/O Devices

• OS also allows certain IPC to be modeled as communication with an abstract I/O device
  - Example: inter-process pipes, mailslots, network sockets
  - This explains why ReadFile is so universal

• Our main focus here is on file I/O, but similar principles apply to other types of devices
  - Just reading files is simple; however, achieving decent speed and parallelizing computation is more challenging

• Before solving this problem, we start with a general background on files and APIs
  - Homework #3 requires multi-CPU searching of Wikipedia for user-specified substrings
Background on Files

• Just like RAM, a file is a sequence of bytes
• Supports 3 main operations: read, write, and seek
• *File pointer* specifies the current position within the file
  – Read/write operations proceed from that location forward
• Example: test.txt written in notepad:
  – Byte contents give by hex viewer (e.g., HxD)

| 54 68 69 73 20 69 73 20 61 20 74 65 78 74 20 66 69 6C 65 2E 0D 0A 53 65 63 6F 6E 64 20 6C 69 6E 65 2E | This is a text file. Second line. |

• What is the ASCII table?
  – Why is there 0xD and 0xA in the file?
Two modes of file I/O: text and binary
- Must be requested when you open the file

Binary means disk contents are an exact copy of the RAM buffer that is written and vice versa

Text means there is some library (wrapper) between the application and OS that applies certain “magic” translation before your program sees the data
- For fopen/fprintf, this involves \r\n → \n, terminating the read at Ctrl-Z markers (ASCII code 26), and certain multi-byte to wide char mapping based on the locale

Note: text files can be always read in binary mode, while the opposite is not true
### Background on Files

- **Example:** *binary mode* reads the file as is:

  - 54 68 69 73 20 69 73 20 61 20 74 65 78 74 20 66 69 73 20 73 20 74 69 66 69 6C 65 20 6C 69 6E 65 2E

  - while *text mode* removes \r

    - 54 68 69 73 20 69 73 20 61 20 74 65 78 74 20 66 69 73 20 73 20 74 69 66 69 6C 65 20 6C 69 6E 65 2E

- If the file is tweaked before it reaches your program, lots of confusing things may happen
  - E.g., file size 100,050 bytes, but your buffer gets only 99,800

- Since text-mode processing does usually unwanted things to the file and is much slower than binary mode, it is not recommended (see later for benchmarks)
Number representation can be ASCII or native
- ASCII is similar to what you see on the screen
- Native is identical to how numbers are stored in RAM

Example:

```
int x = 0x11223344;
```

Native version: 44 33 22 11
Decimal ASCII version of x, i.e., string “287454020”

ASCII output depends on how the numbers are written (e.g., decimal, hex) and the separator between them
- Conversion to/from ASCII is usually slow
- Format inefficient in terms of storage

APIs that read raw buffers are usually native
- Those that attempt to read individual variables are ASCII
• Suppose we read an integer natively from the beginning of this file

```c
int x;
ReadFile (&x, sizeof(int));
```

- What is the value of x?
- Equivalent versions →

• How to write contents of some class natively to disk?
  - If it has no pointers, then it’s trivial

```c
class MyClass {
    double a;
    uint64 b;
};
MyClass mc;
mc.a = 3.1415;
mc.b = 0x55;
WriteFile (... , &mc, sizeof(MyClass), ...);
```
Background on Files

• How to store pointers, e.g., a linked list or binary tree?

```cpp
class LinkedListElem {
    int val;
    LinkedListElem *next;
};
```

```cpp
class TreeElem {
    int val;
    TreeElem *left, *right;
};
```

• Data structure must first be converted to an array
  - Hierarchical structure must be flattened

```cpp
int valArray = new int [LinkedList.size()];
// traverse the list, copy into valArray
WriteFile (... , valArray,
    sizeof(int) * LinkedList.size() , ...);
```

```cpp
class TreeElem2 {
    int val;
    int left, right; // offsets
};
TreeElem2 *arr = new
    TreeElem2 [tree.size()];
```

```
val = 55
left = 1
right = 2
```

```
val = 22
left = 3
right = 0
```

```
val = 77
left = 4
right = 5
```

```
val = 14
left = 0
right = 0
```

```
val = 65
left = 0
right = 0
```

```
val = 90
left = 0
right = 0
```
Background on Files

• In fact, trees stored as arrays in RAM are often much faster than pointer-based trees
  – Main drawback: difficult to deal with fragmentation

• Further compaction: 2 bits to store # of children
  – Suppose 00 = none, 01 = left, 10 = right, 11 = both

<table>
<thead>
<tr>
<th>val</th>
<th>bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>3</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>77</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>90</td>
<td>0</td>
</tr>
</tbody>
</table>

• Conversion from random-access (RAM) structures to sequential arrays is called **serialization**
  – Similar to serial transmission over COM ports or networks
Background on Files

- Asking the kernel for chunk of data
  - How large should the chunk be?
- Clearly not too small, otherwise many kernel-mode transitions, which are costly
- Some wrapper libraries (FILE and STL streams) have yet another buffer to avoid kernel-mode switching
  - Also needed if they perform text-mode pre-processing
- OS buffering can be disabled
  - Disk driver directly DMAs data into your program’s buffer
  - Caveat: buffer size must be a multiple of sector size (512 bytes)
APIs

• CreateFile is the most flexible and high-performance method of doing I/O
  - Treats the memory as a sequence of bytes
  - Operates in binary mode and gives you the native representation of RAM data structures

• Read MSDN about access (read, write, both), sharing, and disposition (e.g., open existing, create new)

• The flag field sets the attributes (e.g., hidden, encrypted, read-only, archived, system)
  - Also can be used to disable OS buffering (FILE_FLAG_NO_BUFFERING) or enable overlapped operation (FILE_FLAG_OVERLAPPED)
APIs

• Some functions take two DWORDs instead of one uint64
  - How to convert?

// combining DWORDs into uint64
DWORD high, low = GetFileSize (h, &high);
uint64 size = ((uint64)high << 32) + low;

// splitting a uint64 into DWORDs
high = size >> 32;
low = size & ((DWORD) -1);

• Overlapped I/O allows multiple outstanding requests

Note: each pending request must have its own struct ol

char buf [BUF_SIZE];
DWORD bytes;

// read a whole chunk
if (ReadFile (hFile, buf, BUF_SIZE,
    &bytes, NULL) == 0) {
    if (GetLastError () != ERROR_HANDLE_EOF) {
        // handle error
        exit (-1);
    }
    reachedEof = true;
}
else if (bytes < BUF_SIZE)
    reachedEof = true;
printf (“Obtained %d bytes, EOF = %d
”,
    bytes, reachedEof);

DWORD low = GetFileSize (HANDLE hFile,
    LPDWORD high);

DWORD WINAPI SetFilePointer(
    __in HANDLE hFile,
    __in LONG lDistanceToMove,
    __inout_opt PLONG lpDistanceToMoveHigh,
    __in DWORD dwMoveMethod );

OVERLAPPED ol;
memset (&ol, 0, sizeof (OVERLAPPED));
ol.hEvent = CreateEvent (NULL, false, false, NULL);
ReadFile (hFile, buf, len, NULL, &ol);
// if error == ERROR_IO_PENDING, continue
WaitForSingleObject (ol.hEvent, INFINITE);
GetOverlappedResult (hFile, &ol, &bytesRead, false);

DWORD WINAPI SetFilePointer(
    __in HANDLE hFile,
    __in LONG lDistanceToMove,
    __inout_opt PLONG lpDistanceToMoveHigh,
    __in DWORD dwMoveMethod );
• The FILE stream is the classical C-style library
  – Portable to Unix and most other OSes

char buf [BUF_SIZE];
// open for reading in binary mode
FILE *f = fopen ("test.txt", "rb");
if (f == NULL) {
    printf ("Error %d opening file\n", errno);
    exit (-1);
}
// read up to one full buffer
// native representation
int bytesRead = fread (buf, 1, BUF_SIZE, f);
fclose (f);

FILE *f = fopen ("test.txt", "rb");
// seek to the end
_fseeki64 (f, 0, SEEK_END);
// get current position
uint64 fileSize = __ftelli64 (f);
// return to beginning
_fseeki64 (f, 0, SEEK_SET);
printf ("file size %I64u\n", fileSize);

int a = 5;
double b = 10;
// open for writing in binary mode
FILE *f = fopen ("test.txt", "wb");
// ASCII representation
fprintf (f, "a = %d, b = %f\n", a, b);
fclose (f);

int a;
double b;
// ASCII decoding of numbers
int ret = fscanf (f, "%d %f", &a, &b);
if (ret == 0 || ret == EOF)
    printf ("Hit error or EOF\n");
else
    printf ("Obtained %d, %f\n", a, b);
// %s gets one word and NULL terminates it
// note: potential buffer overflow
fscanf (f, "%s", buf);
// recommended to specify buf length
fscanf (f, "%32s", buf);
**APIs**

- If an entire line is needed, a faster alternative to fscanf is fgets()
- STL streams are similar

```c
char buf [BUF_SIZE];
FILE *f = fopen ("test.txt", "rb");
while (!feof (f)) {
    // read one line at a time
    if (fgets (buf, BUF_SIZE, f) == NULL)
        break; // EOF or error
    printf ("Line '%s' has %d bytes\n", buf, strlen(buf));
}
fclose (f);
```

- Q: using Windows APIs, how to print contents of a text file?

```c
// assume file is small and fits in RAM
// allocate the buffer
char *buf = new char [fileSize + 1];
ReadFile (... , buf, fileSize, &bytes, ...);
// TODO: error checks
buf[bytes] = NULL;
printf ("%s\n", buf);
```
Performance

- Dual RAID controllers, each with 12 disks in RAID-5
  - Speed given in MB/s, CPU utilization = fraction of 16 cores
    - Text mode
      - ifs >> s: 1.8/12
      - fscanf (f, “%s”, buf): 6/19
      - fgets (buf, BUF_SIZE, f): 26/50
      - ifs.read w/32MB buffer: 90
      - fread w/32MB buffer: 90/144
      - ReadFile w/32MB buffer
      - ReadFile + no OS buffering
      - ReadFile + no buf + overlapped: 2500
    - Binary mode
      - Debug
      - Release
      - CPU utilization: 10%/9%/7%/10%/11%/10%/11%
- Modern PCI-e 4.0 m.2 drives in RAID
  - Up to 30 GB/s