CS 33

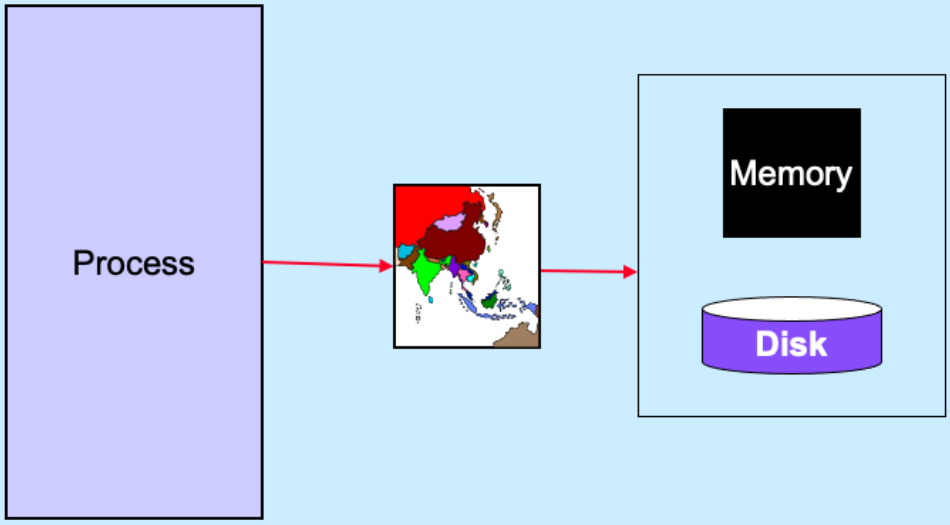
Virtual Memory 2

OS Role in Virtual Memory

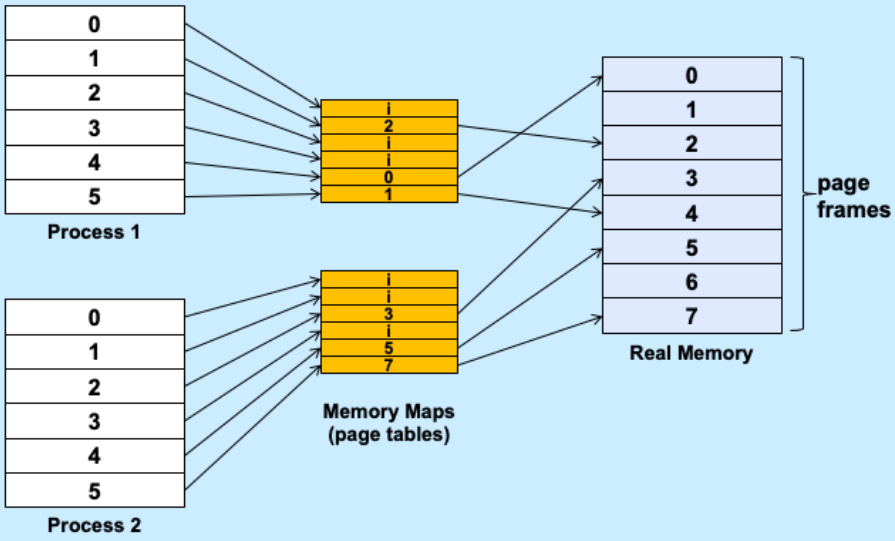
- Memory is like a cache
  - quick access if what's wanted is mapped via page table
  - slow if not — OS assistance required
- OS
  - make sure what's needed is mapped in
  - make sure what's no longer needed is not mapped in

Why is virtual memory used?

More VM than RM

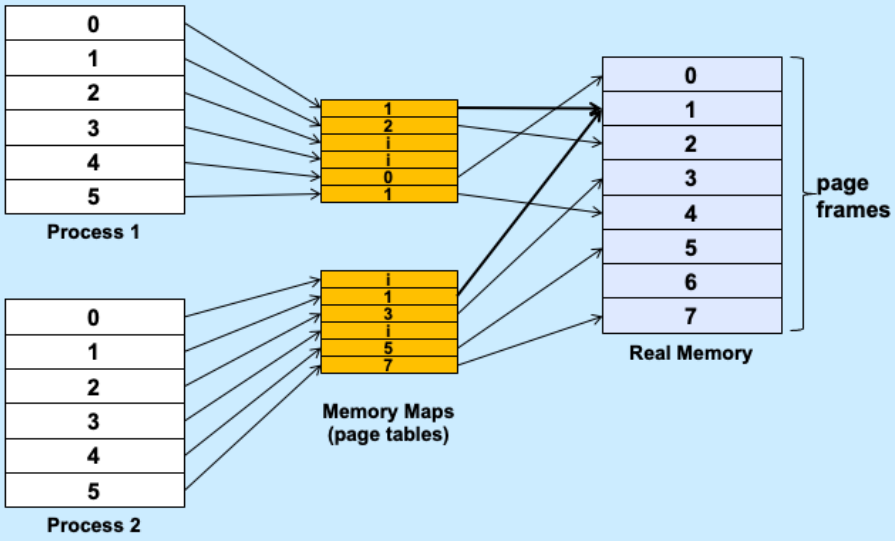


Isolation

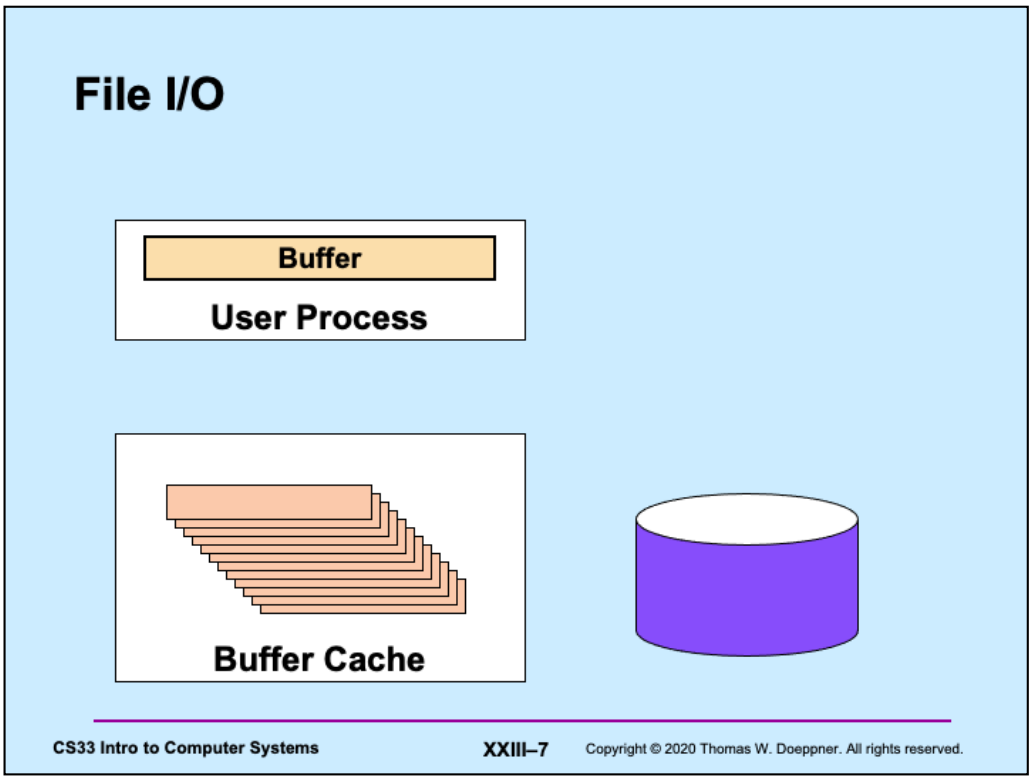


Virtual Memory

Sharing



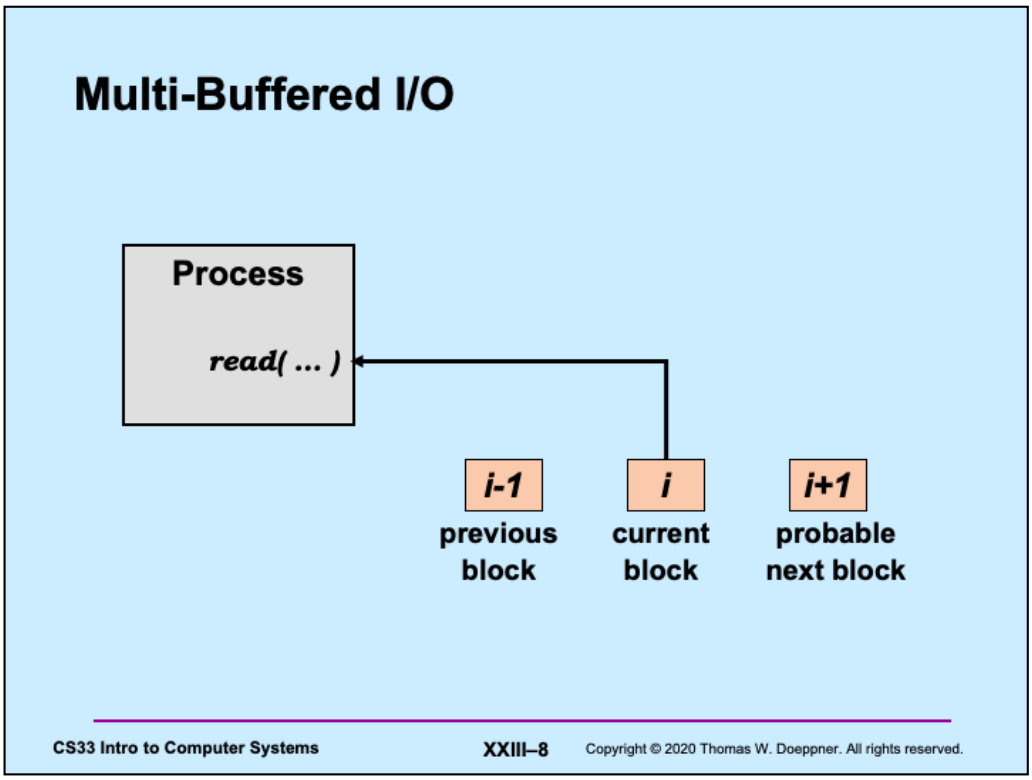
Virtual Memory



File I/O in Unix, and in most operating systems, is not done directly to the disk drive, but through intermediary buffers, known as the buffer cache, in the operating system's address space. This cache has two primary functions. The first, and most important, is to make possible concurrent I/O and computation within a Unix process. The second is to insulate the user from physical disk-block boundaries.

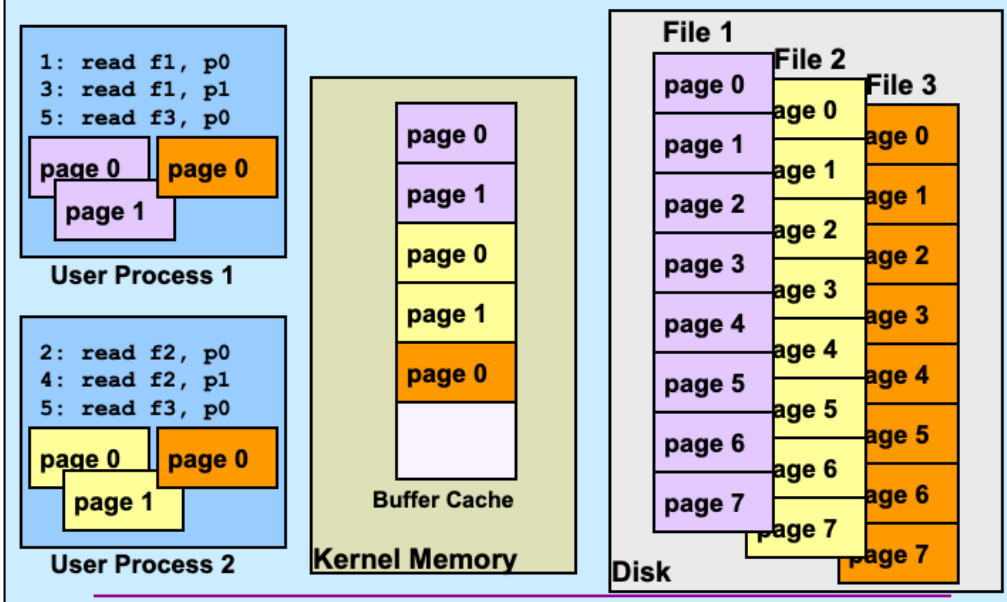
From a user process's point of view, I/O is *synchronous*. By this we mean that when the I/O system call returns, the system no longer needs the user-supplied buffer. For example, after a write system call, the data in the user buffer has either been transmitted to the device or copied to a kernel buffer — the user can now scribble over the buffer without affecting the data transfer. Because of this synchronization, from a user process's point of view, no more than one I/O operation can be in progress at a time.

The buffer cache provides a kernel implementation of multibuffered I/O, and thus concurrent I/O and computation are made possible.

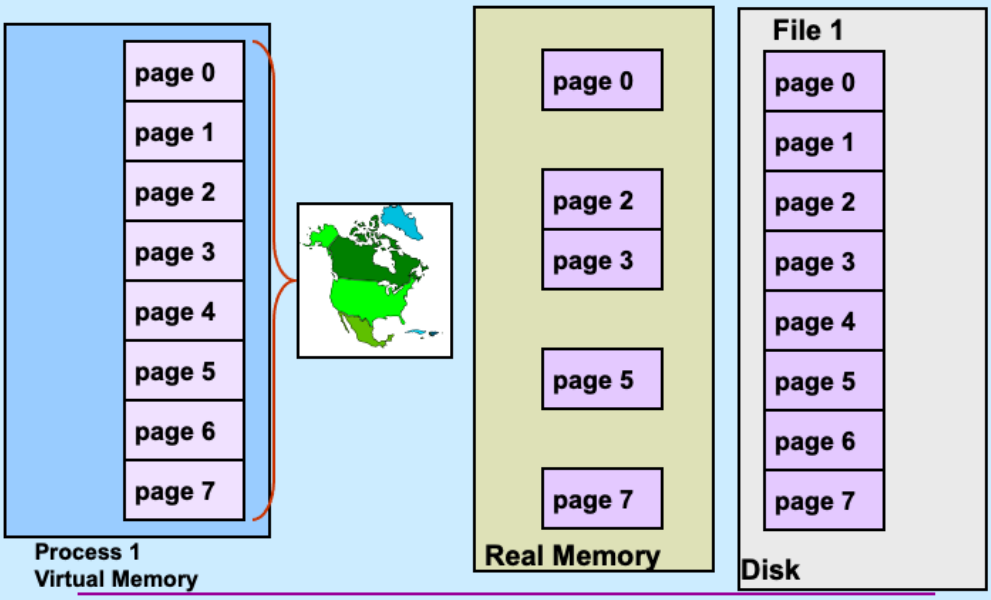


The use of *read-aheads* and *write-behinds* makes possible concurrent I/O and computation: if the block currently being fetched is block *i* and the previous block fetched was block *i-1*, then block *i+1* is also fetched. Modified blocks are normally written out not synchronously but instead sometime after they were modified, asynchronously.

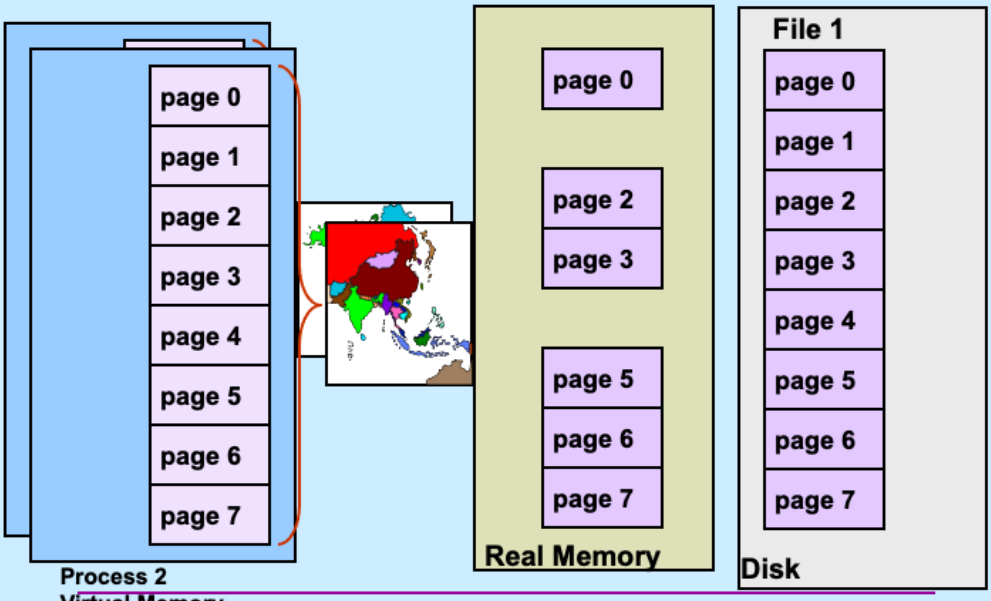
Traditional I/O



Mapped File I/O



Multi-Process Mapped File I/O



Mapped Files

- **Traditional File I/O**

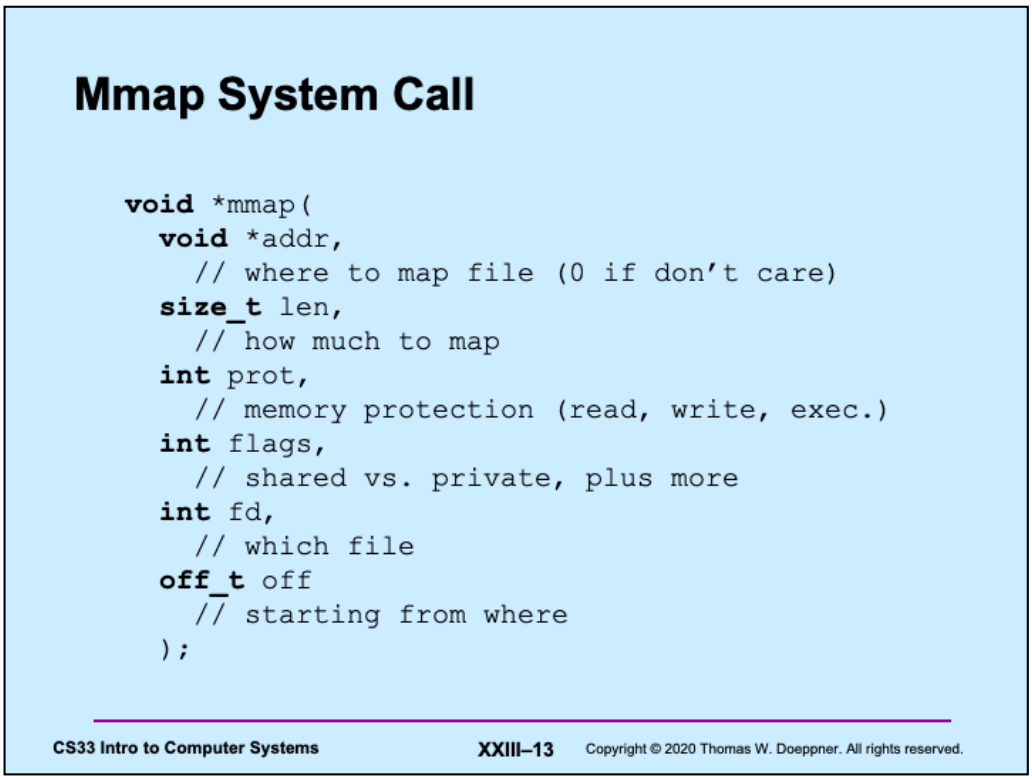
```
char buf[BigEnough];
fd = open(file, O_RDWR);
for (i=0; i<n_recs; i++) {
    read(fd, buf, sizeof(buf));
    use(buf);
}
```

- **Mapped File I/O**

```
record_t *MappedFile;
fd = open(file, O_RDWR);
MappedFile = mmap(... , fd, ...);
for (i=0; i<n_recs; i++)
    use(MappedFile[i]);
```


Traditional I/O involves explicit calls to read and write, which in turn means that data is accessed via a buffer; in fact, two buffers are usually employed: data is transferred between a user buffer and a kernel buffer, and between the kernel buffer and the I/O device.

An alternative approach is to *map* a file into a process's address space: the file provides the data for a portion of the address space and the kernel's virtual-memory system is responsible for the I/O. A major benefit of this approach is that data is transferred directly from the device to where the user needs it; there is no need for an extra system buffer.

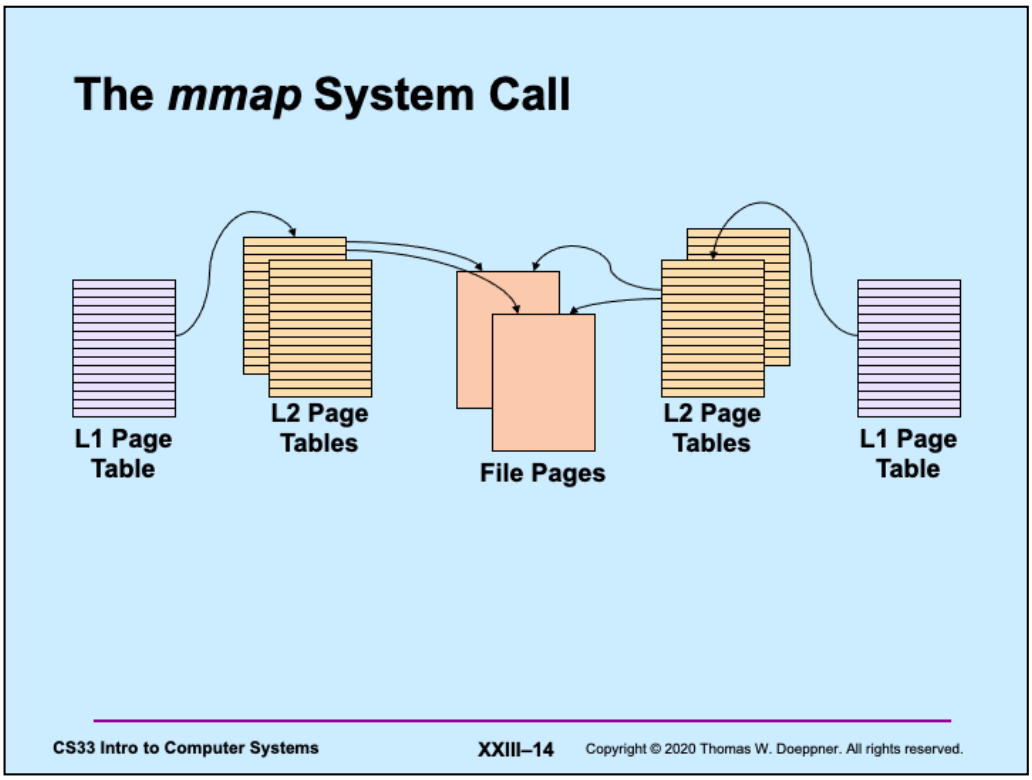


Mmap maps the file given by *fd*, starting at position *off*, for *len* bytes, into the caller's address space starting at location *addr*

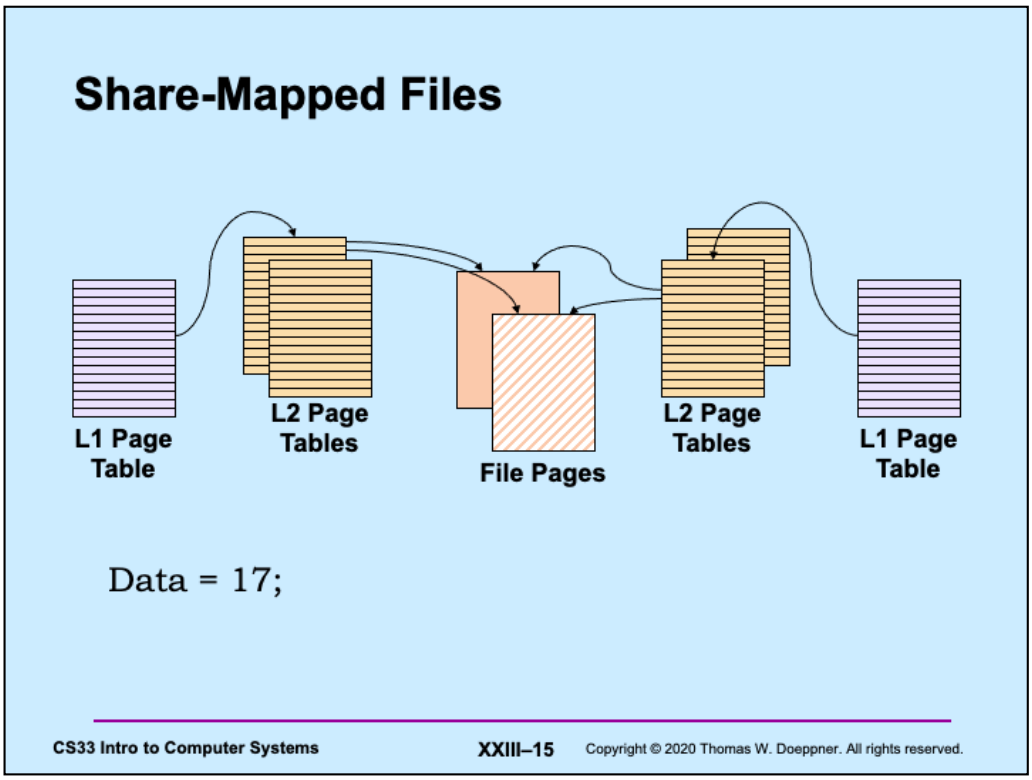
- *len* is rounded up to a multiple of the page size
- *off* must be page-aligned
- if *addr* is zero, the kernel assigns an address
- if *addr* is positive, it is a suggestion to the kernel as to where the mapped file should be located (it usually will be aligned to a page). However, if *flags* includes MAP_FIXED, then *addr* is not modified by the kernel (and if its value is not reasonable, the call fails)
- the call returns the address of the beginning of the mapped file

The flags argument must include either MAP_SHARED or MAP_PRIVATE (but not both). If it's MAP_SHARED, then the mapped portion of the caller's address space contains the current contents of the file; when the mapped portion of the address space is modified by the process, the corresponding portion of the file is modified.

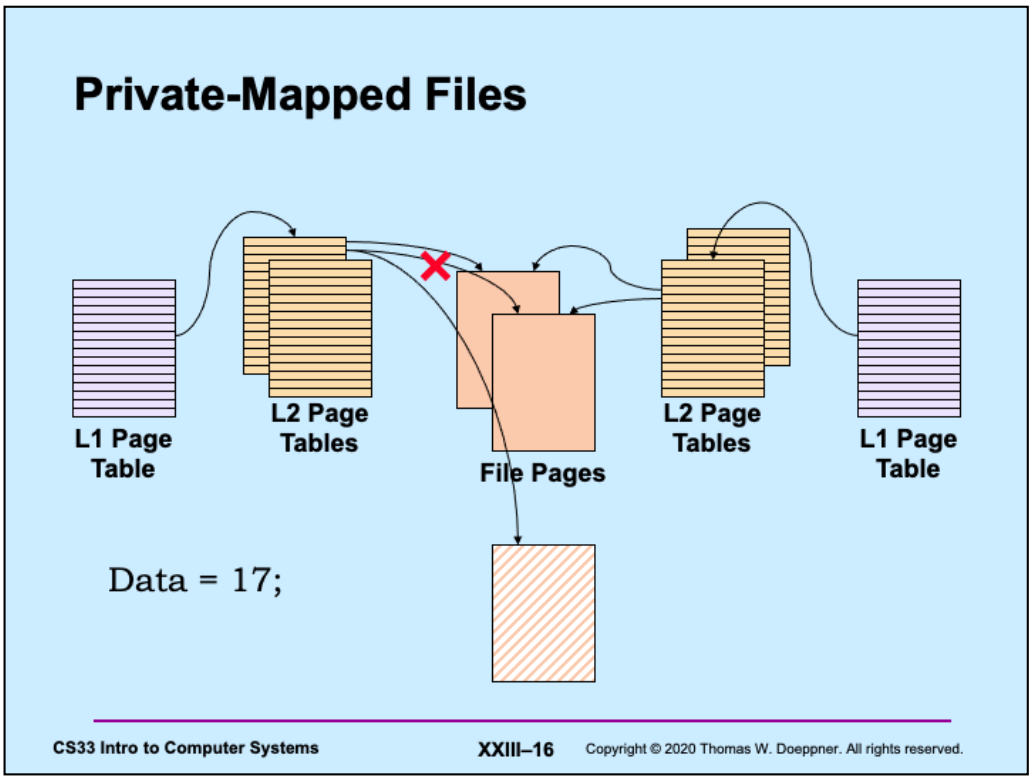
However, if *flags* includes MAP_PRIVATE, then the idea is that the mapped portion of the address space is initialized with the contents of the file, but that changes made to the mapped portion of the address space by the process are private and not written back to the file. The details are a bit complicated: as long as the mapping process does not modify any of the mapped portion of the address space, the pages contained in it contain the current contents of the corresponding pages of the file. However, if the process modifies a page, then that particular page no longer contains the current contents of the corresponding file page, but contains whatever modifications are made to it by the process. These changes are not written back to the file and not shared with any other process that has mapped the file. It's unspecified what the situation is for other pages in the mapped region after one of them is modified. Depending on the implementation, they might continue to contain the current contents of the corresponding pages of the file until they, themselves, are modified. Or they might also be treated as if they'd just been written to and thus no longer be shared with others.



The *mmap* system call maps a file into a process's address space. All processes mapping the same file can share the pages of the file.



There are a couple options for how modifications to mmapped files are dealt with. The most straightforward is the *share* option in which changes to mmapped file pages modify the file and hence the changes are seen by the other processes who have share-mapped the file.



The other option is to *private*-map the file: changes made to mmapped file pages do not modify the file. Instead, when a page of a file is first modified via a private mapping, a copy of just that page is made for the modifying process, but this copy is not seen by other processes, nor does it appear in the file.

In the slide, the process on the left has private-mapped the file. Thus its changes to the mapped portion of the address space are made to a copy of the page being modified.

Example

```
int main( ) {
  int fd;
  dataObject_t *dataObjectp;

  fd = open("file", O_RDWR);
  if ((int)(dataObjectp = (dataObject_t *)mmap(0,
      sizeof(dataObject_t),
      PROT_READ|PROT_WRITE, MAP_SHARED, fd, 0)) == -1) {
    perror("mmap");
    exit(1);
  }

  // dataObjectp points to region of (virtual) memory
  // containing the contents of the file

  ...

}
```


Here we map the contents of a file containing a dataObject_t into the caller's address space, allowing it both read and write access. Note mapping the file into memory does not cause any immediate I/O to take place. The operating system will perform the I/O when necessary, according to its own rules.

fork and mmap

```
int main() {
  int x=1;

  if (fork() == 0) {
    // in child
    x = 2;
    exit(0);
  }
  // in parent
  while (x==1) {
    // will loop forever
  }
  return 0;
}
```

```
int main() {
  int fd = open( ... );
  int *xp = (int *)mmap(...,
      MAP_SHARED, fd, ...);
  xp[0] = 1;
  if (fork() == 0) {
    // in child
    xp[0] = 2;
    exit(0);
  }
  // in parent
  while (xp[0]==1) {
    // will terminate
  }
  return 0;
}
```


When a process calls fork and creates a child, the child's address space is normally a copy of the parent's. Thus changes made by the child to its address space will not be seen in the parent's address space (as shown in the left-hand column). However, if there is a region in the parent's address space that has been mmapped using the MAP_SHARED flag, and subsequently the parent calls fork and creates a child, the mmapped region is not copied but is shared by parent and child. Thus changes to the region made by the child will be seen by the parent (and vice versa).

Putting Together a Program

gcc Steps

1) Compile
   - to start here, supply .c file
   - to stop here: `gcc -S` (produces .s file)
   - if not stopping here, gcc compiles directly into a .o file, bypassing the assembler
2) Assemble
   - to start here, supply .s file
   - to stop here: `gcc -c` (produces .o file)
3) Link
   - to start here, supply .o file

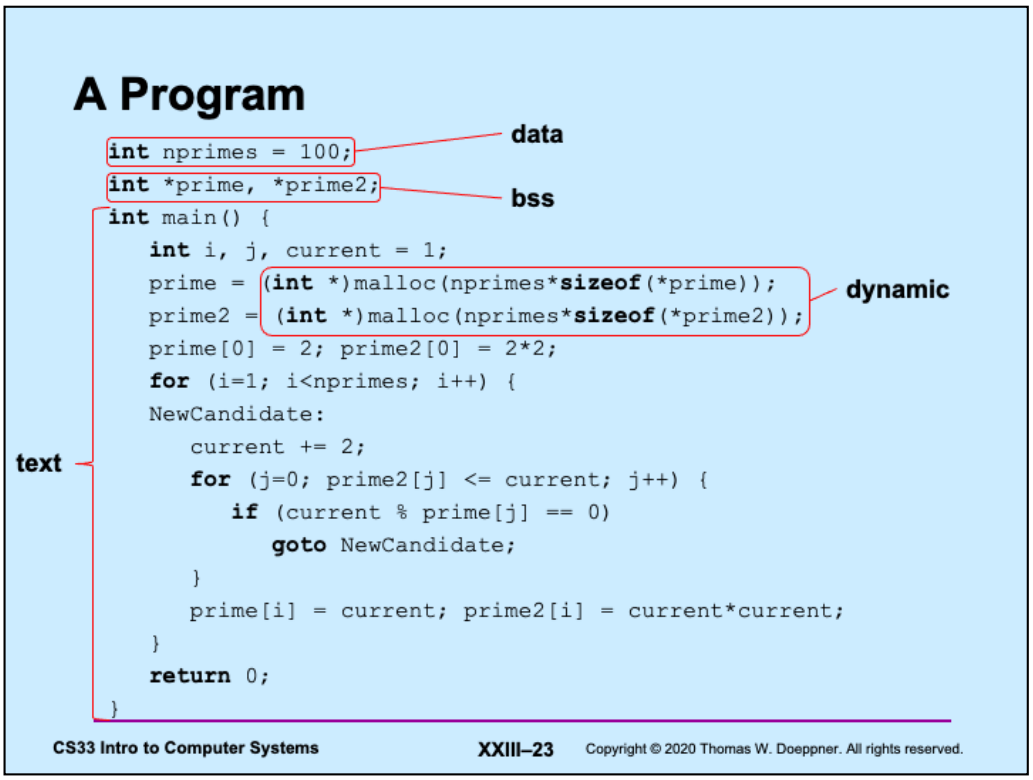
The Linker

- An executable program is one that is ready to be loaded into memory
- The linker (known as ld: /usr/bin/ld) creates such executables from:
  - object files produced by the compiler/assembler
  - collections of object files (known as libraries or archives)
  - and more we'll get to soon ...

The technology described in this and the next few slides was commonplace around 1990 and is known as *static linking*. We discuss static linking first, then later move on to *dynamic linking*, which is commonplace today.

Linker's Job

- Piece together components of program
  - arrange within address space
    - » code (and read-only data) goes into text region
    - » initialized data goes into data region
    - » uninitialized data goes into bss region
- Modify address references, as necessary



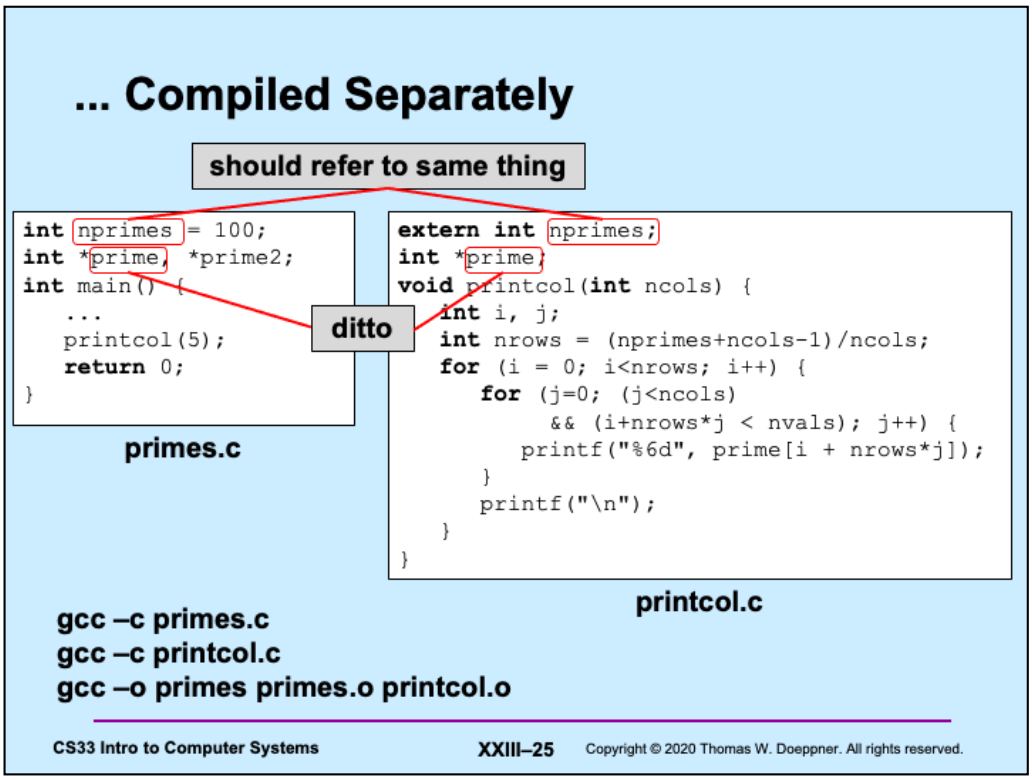
The code is an implementation of the "sieve of Eratosthenes", an early (~200 BCE) algorithm for enumerating prime numbers.

... with Output

```
int nprimes = 100;
int *prime, *prime2;
int main() {
   ...
   printcol(5);
   return 0;
}

void printcol(int ncols) {
   int i, j;
   int nrows = (nprimes+ncols-1)/ncols;
   for (i = 0; i<nrows; i++) {
      for (j=0; (j<ncols) && (i+nrows*j < nvals); j++) {
         printf("%6d", prime[i + nrows*j]);
      }
      printf("\n");
   }
}
```


What this program actually does isn't all that important for our discussion. However, it prints out the vector of prime numbers in multiple columns.

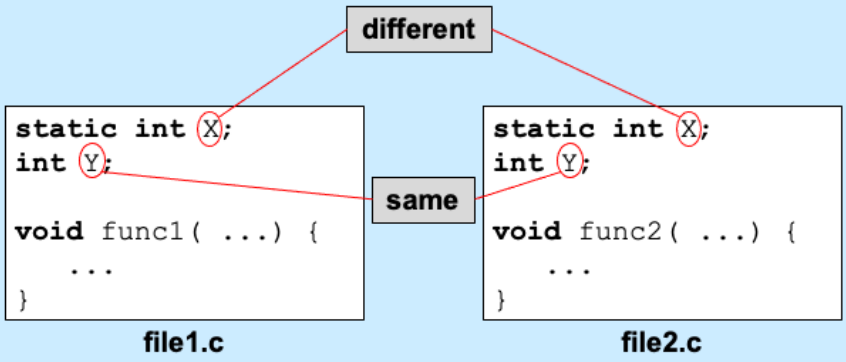


In the first two invocations of gcc, the "-c" flag tells it to compile the C code and produce an object (".o") file, but not to go any further (and thus not to produce an executable program). In the third invocation, gcc invokes the ld (linker) program to combine the two object files into an executable program. As we discuss soon, it will also bring in code (such as printf) from libraries.

Global Variables

- Initialized vs. uninitialized
  - initialized allocated in *data* section
  - uninitialized allocated in *bss* section
    - » implicitly initialized to zero
- File scope vs. program scope
  - *static* global variables known only within file that declares them
    - » two of same name in different files are different
    - » e.g., `static int X;`
  - non-static global variables potentially shared across all files
    - » two of same name in different files are same
    - » e.g., `int X;`

Scope

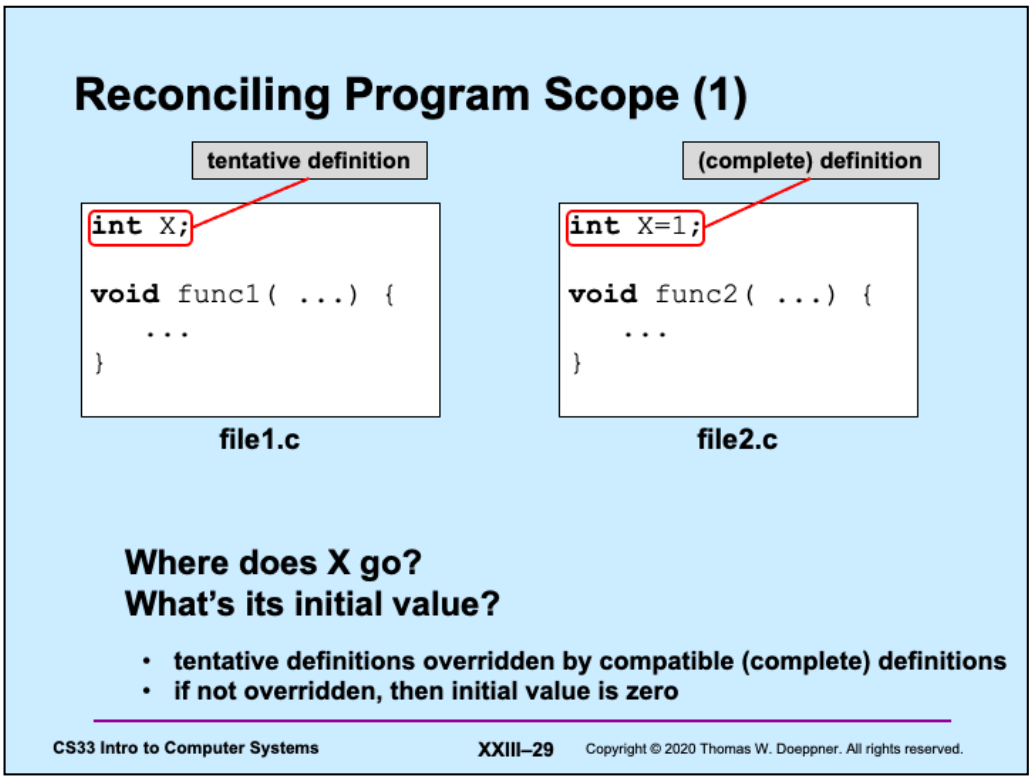


Static Local Variables

```
int *sub1() {
  int var = 1;
  …
  return &var;
  /* amazingly illegal */
}
```

```
int *sub2() {
  static int var = 1;
  …
  return &var;
  /* (amazingly) legal */
}
```


Static local variables have the same scope as other local variables, but their values are retained across calls to the procedures they are declared in. Like global variables, uninitialized static local variables are stored in the BSS section of the address space (and implicitly initialized to zero), initialized static local variables are stored in the data section of the address space.



X goes in the data section and has an initial value of 1. If file2.c did not exist, then X would go in the bss section and have an initial value of 0. Note that the textbook calls tentative definitions “weak definitions” and complete definitions “strong definitions”. This is non-standard terminology and conflicts with the standard use of the term “weak definition".

Reconciling Program Scope (2)

```
int X=2;

void func1( ...) {
    ...
}
```

file1.c

```
int X=1;

void func2( ...) {
    ...
}
```

file2.c

What happens here?

In this case we have conflicting definitions of X — this will be flagged (by the ld program) as an error.

Reconciling Program Scope (3)

```
int X=1;

void func1( ...) {
    ...
}
```

file1.c

```
int X=1;

void func2( ...) {
    ...
}
```

file2.c

Is this ok?

No; it is flagged as an error: only one file may supply an initial value.

Reconciling Program Scope (4)

```
extern int X;

void func1( ...) {
    ...
}
```

file1.c

```
int X=1;

void func2( ...) {
    ...
}
```

file2.c

What's the purpose of "extern"?

The "extern" means that this file will be using X, but it depends on some other file to provide a definition for it, either initialized of uninitialized. If no other file provides a definition, then ld flags an error.

If the "extern" were not there, i.e., if X were declared simply as an "int" in file1.c, then it wouldn't matter if no other file provided a definition for X — X would be allocated in bss with an implicit initial value of 0.

Note: this description of extern is how it is implemented by gcc. The official C99 standard doesn't require this behavior, but merely permits it. It also permits "extern" to be essentially superfluous: its presence may mean the same thing as its absence.

The C11 standard more-or-less agrees with the C99 standard. Moreover, it explicitly allows a declaration of the form "extern int X=1;" (i.e., initialization), which is not allowed by gcc.

For most practical purposes, whatever gcc says is the law ...

Does Location Matter?

```
int main(int argc, char *[]) {
   return(argc);
}
```

```
main:
  pushq %rbp    ; push frame pointer
  movq  %rsp, %rbp     ; set frame pointer to point to new frame
  movl  %edi, %eax     ; put argc into return register (eax)
  movq  %rbp, %rsp     ; restore stack pointer
  popq  %rbp    ; pop stack into frame pointer
  ret          ; return: pops end of stack into rip
```


This rather trivial program references memory via only rsp and rip (rbp is set from rsp). Its code contains no explicit references to memory, i.e., it contains no explicit addresses.

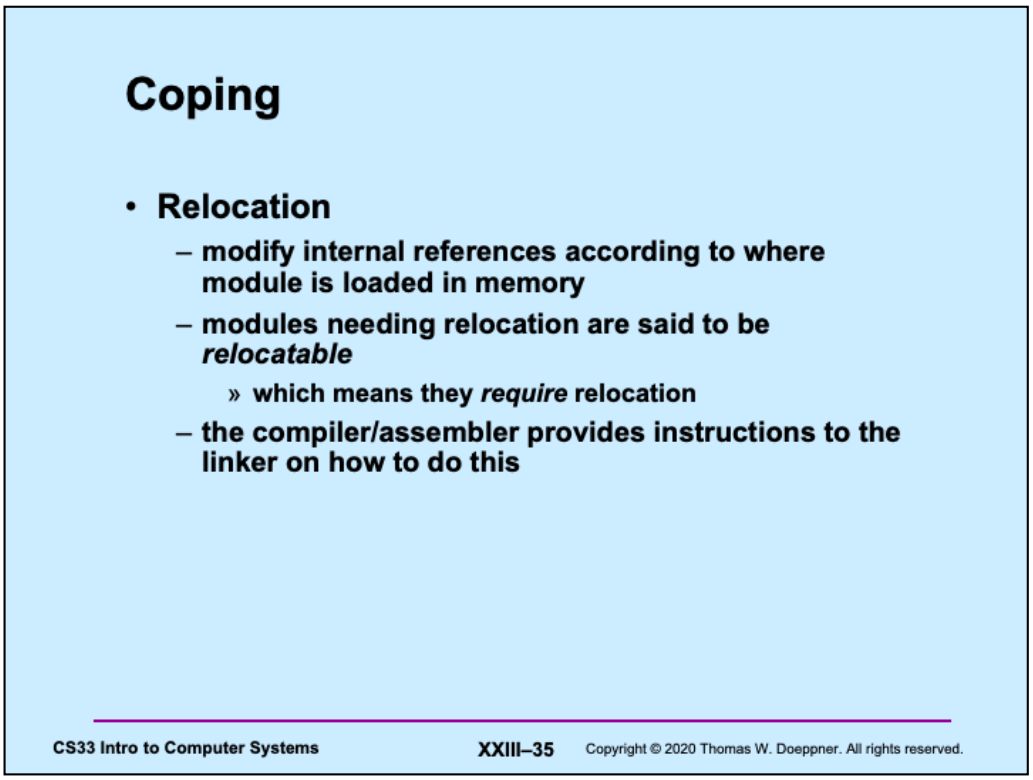
Location Matters …

```
int X=6;
int *aX = &X;

int main() {
    void subr(int);
    int y=*aX;
    subr(y);
    return(0);
}

void subr(int i) {
    printf("i = %d\n", i);
}
```


We don't need to look at the assembler code to see what's different about this program: the machine code produced for it can't simply be copied to an arbitrary location in our computer's memory and executed. The location identified by the name *aX* should contain the address of the location containing *X*. But since the address of *X* will not be known until the program is copied into memory, neither the compiler nor the assembler can initialize *aX* correctly. Similarly, the addresses of *subr* and *printf* are not known until the program is copied into memory — again, neither the compiler nor the assembler would know what addresses to use.



Due to the shortened semester, we won't cover the details of how relocation is done. Linux uses an approach known as ELF (executable and linkable format). Apple's OSX uses a different approach known as Mach-O (for Mach object format – OSX was derived from an OS produced in the 1980s at Carnegie Mellon University known as Mach).