

GNU/Linux

Application Programming

Lecture Handouts

WANG Xiaolin

`wx672ster+linux@gmail.com`

August 25, 2019

Contents

Part I	Getting Started	4
Part II	Shell Basics	6
1	Shell Basics	6
2	Shell Programming	9
Part III	Linux Programming Environment	14
3	C Programming Environment	14
3.1	The Tool Chain	14
3.2	Header Files	15
3.3	Library Files	16
3.4	Error Handling	19
3.5	The Make Utility	19
3.6	Version Control	20
3.7	Manual Pages	21
3.8	A Sample GNU Package	21
4	The Linux Environment	21
5	OS Basics	24
5.1	Hardware	25
5.2	Bootstrapping	27
5.3	Interrupt	28
5.4	System Calls	29
Part IV	Working With Files	33
6	Files	33
7	Directories	40
Part V	Processes and Threads	44





8 Virtual Memory	44
9 Process	53
10 Thread	55
11 Signals	60
 Part VI Interprocess Communication	 63
12 Pipes and FIFOs	64
13 Message Queues	70
14 Semaphores	73
15 Classical IPC Problems	81
15.1 The Dining Philosophers Problem	81
15.2 The Readers-Writers Problem	84
15.3 The Sleeping Barber Problem	84
16 Shared Memory	85
17 Sockets	86
17.1 Stream Sockets	89
17.2 Datagram Sockets	90
17.3 Unix Domain Sockets	90
17.4 Internet Sockets	92

References

- [1] VENKATESH B, ANGRAVE L, et Al. *CS241 System Programming Coursebook*. University of Illinois, 2019.
- [2] MATTHEW N, STONES R. *Beginning linux programming*. 4th ed. John Wiley & Sons, 2008.
- [3] COOPER M. *Advanced Bash Scripting Guide 5.3 Volume 1*. Lulu.com, 2010.
- [4] RAYMOND E S. *The art of Unix programming*. Addison-Wesley, 2003.
- [5] STEVENS W R, RAGO S A. *Advanced programming in the UNIX environment*. Addison-Wesley, 2013.
- [6] LOVE R. *Linux System Programming: Talking Directly to the Kernel and C Library*. O'Reilly Media, Inc., 2007.
- [7] KERRISK M. *The Linux Programming Interface: A Linux and UNIX System Programming Handbook*. No Starch Press, 2010.
- [8] BRYANT R E, O'HALLARON D R. *Computer Systems: A Programmer's Perspective*. 2nd ed. Addison-Wesley, 2010.
- [9] TANENBAUM A S. *Modern Operating Systems*. 4th ed. Prentice Hall Press, 2015.
- [10] Wikipedia. *File Allocation Table* — *Wikipedia, The Free Encyclopedia*. 2015. http://en.wikipedia.org/w/index.php?title=File_Allocation_Table&oldid=661104239.
- [11] Wikipedia. *Inode* — *Wikipedia, The Free Encyclopedia*. 2015. <http://en.wikipedia.org/w/index.php?title=Inode&oldid=647736522>.
- [12] Wikipedia. *Compiler* — *Wikipedia, The Free Encyclopedia*. 2015. <http://en.wikipedia.org/w/index.php?title=Compiler&oldid=661266598>.
- [13] Wikipedia. *Assembly language* — *Wikipedia, The Free Encyclopedia*. 2015. http://en.wikipedia.org/w/index.php?title=Assembly_language&oldid=661185928.
- [14] Wikipedia. *Linker (computing)* — *Wikipedia, The Free Encyclopedia*. 2015. [http://en.wikipedia.org/w/index.php?title=Linker_\(computing\)&oldid=652892136](http://en.wikipedia.org/w/index.php?title=Linker_(computing)&oldid=652892136).
- [15] Wikipedia. *Loader (computing)* — *Wikipedia, The Free Encyclopedia*. 2012. [http://en.wikipedia.org/w/index.php?title=Loader_\(computing\)&oldid=520743198](http://en.wikipedia.org/w/index.php?title=Loader_(computing)&oldid=520743198).
- [16] Wikipedia. *Dynamic linker* — *Wikipedia, The Free Encyclopedia*. 2012. http://en.wikipedia.org/w/index.php?title=Dynamic_linker&oldid=517400345.

[17] LEVINE J. *Linkers and Loaders*. Morgan Kaufmann, 2000.

Course Web Links

 <https://cs6.swfu.edu.cn/moodle>
 https://cs2.swfu.edu.cn/~wx672/lecture_notes/linux-app/slides/
 https://cs2.swfu.edu.cn/~wx672/lecture_notes/linux-app/src/
 <https://cs3.swfu.edu.cn/tech>

/etc/hosts

```
202.203.132.241  cs6.swfu.edu.cn
202.203.132.242  cs2.swfu.edu.cn
202.203.132.245  cs3.swfu.edu.cn
```

System Programming <https://github.com/angrave/SystemProgramming/wiki>

Beej's Guides <http://beej.us/guide/>

BLP4e <http://www.wrox.com/WileyCDA/WroxTitle/productCd-0470147628,descCd-DOWNLOAD.html>

TLPI <http://www.man7.org/tlpi/>

□ Homework

Weekly tech question

1. What was I trying to do?
2. How did I do it? (steps)
3. The expected output? The real output?
4. How did I try to solve it? (steps, books, web links)
5. How many hours did I struggle?

✉ wx672ster+linux@gmail.com

🇬🇧 Preferably in English

📖 in stackoverflow style

OR simply show me the tech questions you asked on any website

💀 Oversimplified programs ahead!

Part I

Getting Started

Linux Commands

Where to find them? /bin, /usr/bin, /usr/local/bin,
~/bin, ...

```
$ echo $PATH
```

How to find them? which, whereis, type

Command not found?

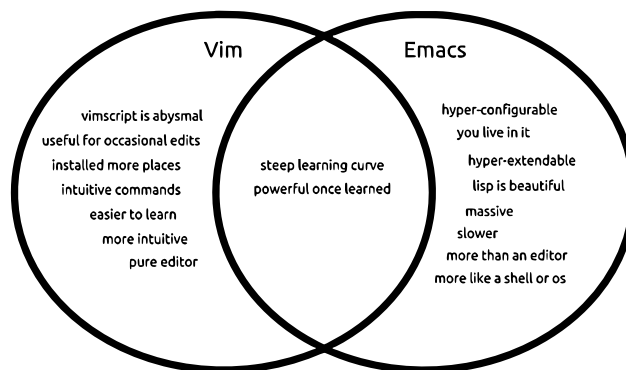
First double check your spelling

Then try:

```
🔍 aptitude search xxx
🔍 apt-cache search xxx
🔍 apt-file search xxx
🔍 sudo apt install packagename
🔍 Google "linux command xxx"
```

Text Editors

 vs. 

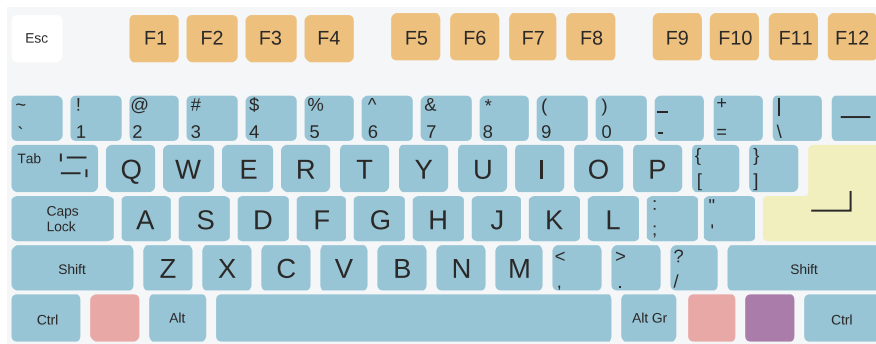


Help Your Editor


Suffix matters

```
$ vim ✗
$ vim hello ✗
$ vim hello.c ✓
$ vim hello.py ✓
$ emacs ✗
$ emacs hello ✗
$ emacsclient hello.c ✓
$ emacsclient hello.py ✓
```

Keyboard



 vimtutor

 `Ctrl` + `h` `t`

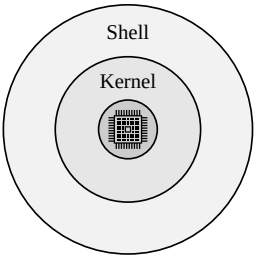
Part II

Shell Basics

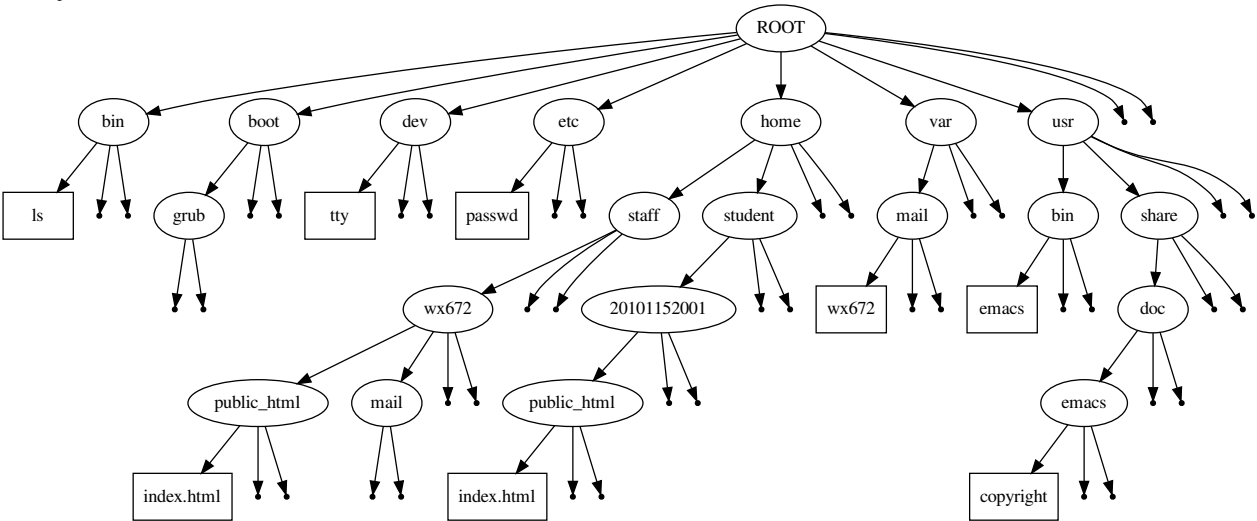
1 Shell Basics

Shell

- ☑ A command line interpreter
- ☑ A programming language



Directory Structure



Todo	How
Where am I?	pwd
What's in it?	ls
Move around?	cd
Disk usage?	du, df
USB drive?	lsblk, mount
New folder?	mkdir

File Operations

Ways to create a file

- ✎ Using an editor (vim, emacs, nano...), or
- \$ cat > filename
- \$ echo "hello, world" > filename
- \$ touch filename

More file operations:

Todo	How	Todo	How
Copy?	cp	Move/Rename?	mv
Delete?	rm	What's it?	file
Link?	ln	Permission?	chmod, chown
Count?	wc	Archive?	tar, gzip, 7z, ...
Sort?	sort, uniq	Search?	find, grep

Redirection

Redirecting output

```
$ ls -l > output.txt
$ ps aux >> output.txt
```

Redirecting input

```
$ more < output.txt
```

Process Operations

Todo	How	Todo	How
Kill?	kill, Ctrl-c	suspend?	Ctrl-z
background?	bg, &	foreground?	fg, jobs
status?	ps, top		

System Info

Todo	How	Todo	How
who?	w, who, whoami	how long?	uptime
software?	apt, aptitude, dpkg	kernel?	uname, lsmod
hardware?	lspci, lsusb, lscpu	memory?	free, lsmem

APT — package management

Todo	How
upgrading?	apt update && apt upgrade
install?	apt install xxx
remove?	apt purge xxx
search?	apt search xxx
details?	apt show xxx
friendly UI?	aptitude

CLI Shortcuts

Ctrl + a :	beginning of line	Ctrl + e :	end of line
Ctrl + f :	forward	Ctrl + b :	backward
Ctrl + n :	next	Ctrl + p :	previous
Ctrl + r :	reverse search	Ctrl + u :	cut to beginning
Ctrl + k :	kill (cut to end)	Ctrl + y :	yank (paste)
Ctrl + d :	delete a character	Tab :	completion

Tmux

Ctrl + a c :	create window	Ctrl + a Ctrl + a :	switch window
Ctrl + a n :	next window	Ctrl + a p :	previous window
Ctrl + a - :	split window	Ctrl + a l :	split window
Ctrl + a j :	go down	Ctrl + a k :	go up
Ctrl + a l :	go right	Ctrl + a h :	go left

Understanding “ls -l”

```
-rw----- 1 sam sam 57 Apr 17 1998 weather.txt
drwxr-xr-x 6 sam sam 102 Oct 9 1999 web_page
-rw-rw-r-- 1 sam sam 7648 Feb 11 20:41 web_site.tar
-rw----- 1 sam sam 574 Dec 16 1998 file.txt
```

File Name
Modification Time
Size (in bytes)
Group
Owner
Number of hard links
File Permissions
File types

d - directory
- - regular file
l - soft link
c - character device
b - block device
s - socket
p - named pipe (FIFO)

9-bit permission

```
7 5 5
111 101 101
rwx r-x r-x
```

Other
Group
User

```
$ chmod 755 foo      $ chmod 644 foo
$ chmod 000 foo      $ chmod 777 foo
$ chmod a-r foo      $ chmod u+x foo
$ chmod g+w foo      $ chmod go=rx foo
```

Wildcard Expansion

Character	Meaning	Example
?	any one	\$ ls ????.txt
*	zero or more	\$ ls *.c
[]	or	\$ ls *. [ch]
{ }	and	\$ ls *. {c,h,cpp}

Example

```
$ touch {2,3,4,234}. {jpg,png} && ls
```

output:

2.jpg	234.jpg	3.jpg	4.jpg
2.png	234.png	3.png	4.png

```
$ rm [234].jpg      $ rm ?.jpg
$ rm {2,3,4,234}.jpg $ rm ?.*
$ rm 2*             $ rm *
```

Everything Is A File

```
$ cat /dev/null > /var/log/messages # empty a file
$ : > /var/log/messages # no new process
$ ls > /dev/null
$ dd if=/dev/zero of=/tmp/clean bs=1k count=1k
$ dd if=/dev/urandom of=/tmp/random bs=1k count=1k
```

/proc

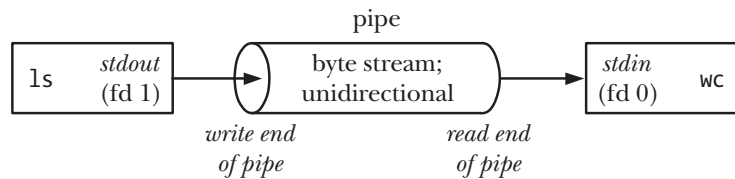
Allow higher-level access to driver and kernel information

```
$ cat /proc/cpuinfo
$ cat /proc/meminfo
$ cat /proc/version
$ cat /proc/1/status
# echo 100000 > /proc/sys/kernel/pid_max
```


Pipe

Chain processes together

```
$ ls | wc -l
```



Unnamed pipe

```
$ unicode skull | head -1 | cut -f1 -d' ' | sm -
```

Named pipe

1. `$ mkfifo mypipe`
 2. `$ gzip -9 -c < mypipe > out.gz`
 3. `$ cat file > mypipe`
- https://en.wikipedia.org/wiki/Named_pipe

2 Shell Programming

\$ — Give Me The Value Of ...

\$var Give me the value of variable “var”

\$(echo hello) Give me the value (output) of command “echo hello”

\$((1+1)) Give me the value (result) of “1+1”

\$\$ Give me the value of special variable “\$”

\$? Give me the value of special variable “?”

\$0 Give me the value of special variable “0”

@ Give me the value of special variable “@”

Variables

```
$ a=8; b=2
$ a=a+5; a=$a+5 ☹️
$ let a=a+5; let a+=5 😊
$ let b=b+a; let b+=a 😊
$ echo a; echo $a
$ (( a=5, b=6, a+=b )) 😊
$ (( b=a<5?8:9 )) 😊
$ r=$(( RANDOM%100 )) 😊
$ echo "$a" # partial quoting
$ echo '$a' # full quoting
$ a=$(ls -l); echo $a; echo "$a"
$ a=hello; b=world; let a+=b ☹️
```

Positional Parameters

\$0, \$1, \$2, ..., \$@, \$#

```

1  #!/bin/bash
2
3  echo "You said:"
4
5  echo -e "\t$t@"
6  echo
7  echo -e "\targc = $# "
8  echo -e "\targv[0] = $0 "
9
10 i=1
11 for arg in $@; do
12     echo -e "\targv[$i] = $arg"
13     let i++
14 done

```

```

1  #include <stdio.h>
2
3  int main(int argc, char *argv[])
4  {
5      int i;
6      printf("You said:\n\t");
7
8      for(i=1; i<argc; i++)
9          printf("%s ",argv[i]);
10
11     printf("\n\n\targc = %d\n", argc);
12
13     for(i=0; i<argc; i++)
14         printf("\targv[%d] = %s\n",i,argv[i]);
15
16     return 0;
17 }
18
19 /* Local Variables: */
20 /* compile-command: "gcc -Wall -Wextra isay.c -o
   ↪ isay" */
21 /* End: */

```

Parameter Substitution

Default value

```

$ echo ${s:=abc}
$ echo ${s:=xyz}

$ echo ${v:-8}
$ echo ${v:-10}

```

Example

```

1  #!/bin/bash
2
3  echo -n Hello, ${1:-world}
4  echo !

```



Re-write it in C

```

$ sudo apt install abs-guide
• file:///usr/share/doc/abs-guide/html/parameter-substitution.html

```

Parameter Substitution

Substring removal

```
$ for f in *.pbm; do ppm2tiff $f ${f%.pbm}.tif; done
```

Substring replacement

```
$ for f in *.jpg; do mv $f ${f%.jpg}.JPG; done
```

Environmental Variables

Each process has an environment

\$PATH	\$PWD	\$HOME	\$UID	\$USER
\$GROUPS	\$SHELL	\$TERM	\$DISPLAY	\$TEMP
\$HOSTNAME	\$HOSTTYPE	\$IFS	\$EDITOR	\$BROWSER
\$HISTSIZE	\$FUNCNAME	\$TMOUT	...	

```

$ export HISTSIZE=2000
$ export BROWSER='/usr/bin/x-www-browser'

```

```
$ export EDITOR='vim'
$ export ALTERNATE_EDITOR="vim"
$ export PDFVIEWER='/usr/bin/zathura'

$ env
$ declare
```

Tests

```
$ (( 5 < 6 )) && echo should be
$ [[ 1 < 2 ]] && echo of course
$ [[ $a -lt $b ]] && echo yes || echo no
$ if [[ $a -lt $b ]]; then echo yes; else echo no; fi
$ if test $a -lt $b; then echo of course; fi
$ if a = 5; then echo a=$a; fi # whitespace matters ✗
$ if a=5; then echo a=$a; fi ☹️
$ if test a=5; then echo a=$a; fi ☹️
$ if test a = 5; then echo a=$a; fi ☹️
$ if test $a = 5; then echo a=$a; fi ✔️
$ test $a = 5 && echo a=$a ✔️
$ [[ $a = 5 ]] && echo a=$a ✔️
$ [[ cmp a b ]] && echo same file ✗
$ if test cmp a b; then echo same file; fi ✗
$ if cmp a b; then echo same file; fi ✔️
$ [[ -f ~/.bash_aliases ]] && . ~/.bash_aliases
$ [[ -x /usr/bin/xterm ]] && /usr/bin/xterm -e tmux &
$ [[ "$pass" != "$MYPASS" ]] && echo 'Wrong password!' && exit 1
$ help test
```

```
1 #!/bin/bash
2
3 words=$@
4 string=linux
5 if echo "$words" | grep -q "$string"
6 then
7     echo "$string found in $words"
8 else
9     echo "$string not found in $words"
10 fi
```

Loops

```
for ARG in LIST; do COMMAND(s); done
$ for i in 1 2 3; do echo -n "$i "; done
$ for i in {1..10}; do echo $i; done
$ for i in $(seq 10); do echo $i; done
$ for ((i=1; i<=10; i++)); do echo $i; done
$ for ((i=1, j=1; i<=10; i++, j++)); do
    echo $i-$j ☹️
    echo $((i-j)) 😊
done
$ for ((i=1; i<=10; i++)) { echo $i; } # C style
$ for i in hello world; do echo -n "$i "; done
```

Loops

```
while CONDITION; do COMMAND(s); done
$ a=0; while [[ a < 10 ]]; do echo $a; ((a++)); done ☹️
$ while [[ $a < 10 ]]; do echo $a; ((a++)); done ☹️
$ while [[ $a -lt 10 ]]; do echo $a; ((a++)); done ✔️
$ while [ $a -lt 10 ]; do echo $a; ((a++)); done ✔️
$ while (( a < 10 )); do echo $a; ((a++)); done ✔️
```

```

$ until (( a = 10 )); do echo $a; ((a++)); done ❌
$ until (( a == 10 )); do echo $a; ((a++)); done ✅
$ while read n; do n2 $n; done
$ while read n; do n2 $n; done < datafile
$ until (( n == 0 )); do read n; n2 $n; done

```

case

```

1  #!/bin/bash
2
3  [ -z "$1" ] && echo "Usage: `basename $0` [dhb]<number>" && exit 0;
4
5  case "$1" in
6      [dD]*)
7          NUM=$(echo $1 | cut -b 2-)
8          printf "\tDec\tHex\tBin\n"
9          printf "\t%d\t0x%02X\t%s\n" $NUM $NUM $(bc <<< "obase=2;$NUM")
10         ;;
11     [hH]*)
12         NUM=$(echo $1 | cut -b 2-)
13         NUM=$(echo $NUM | tr [:lower:] [:upper:])
14         printf "\tHex\t\tDec\t\tBin\n"
15         printf "\t0x%s\t\t%s\t\t%s\n" $NUM $(bc <<< "ibase=16;obase=A;$NUM") \
16             $(bc <<< "ibase=16;obase=2;$NUM")
17         ;;
18     O[xX]*)
19         NUM=$(echo $1 | cut -b 3-)
20         NUM=$(echo $NUM | tr [:lower:] [:upper:])
21         printf "\tHex\t\tDec\t\tBin\n"
22         printf "\t0x%s\t\t%s\t\t%s\n" $NUM $(bc <<< "ibase=16;obase=A;$NUM") \
23             $(bc <<< "ibase=16;obase=2;$NUM")
24         ;;
25     [bB]*)
26         NUM=$(echo $1 | cut -b 2-)
27         printf "\tBin\t\tHex\t\tDec\n"
28         printf "\t%s\t\t0x%s\t\t%s\n" $NUM $(bc <<< "ibase=2;obase=10000;$NUM") \
29             $(bc <<< "ibase=2;obase=1010;$NUM")
30         ;;
31     *)
32         printf "Dec\tHex\tBin\n"
33         printf "%d\t0x%08X\t%08d\n" $1 $1 $(bc <<< "obase=2;$1")
34         ;;
35 esac

```

select

```

1  #!/bin/bash
2
3  PS3='Your favorite OS? '
4
5  select OS in "Linux" "Mac OSX" "Windows"
6  do
7      [[ "$OS" = "Linux" ]] && echo wise guy.
8      [[ "$OS" = "Mac OSX" ]] && echo rich guy.
9      [[ "$OS" = "Windows" ]] && echo patient guy.
10     break
11 done

```

Functions

```

1  #!/bin/bash
2
3  function screencapture(){
4      ffmpeg -f x11grab -s 1920x1080 -r 30 -i :0.0 \
5          -c:v libx264 -crf 0 -preset ultrafast screen.mkv
6  }
7
8  w2pdf(){
9      libreoffice --convert-to pdf:writer_pdf_Export "$1"
10 }
11
12 rfc(){
13     [[ -n "$1" ]] || {
14         cat <<EOF
15         rfc - Command line RFC viewer
16         Usage: rfc <index>
17     EOF
18         return 1
19     }
20     find /usr/share/doc/RFC/ -type f -iname "rfc$1.*" | xargs less
21 }

```

Array

```

1  #!/bin/bash
2
3  IMGDIR="$HOME/Pics/2009Summer/wallpapers/2009summer-1280x768"
4
5  files=( $IMGDIR/*.jpg )
6
7  # get the length of array ${files[@]}
8  n=${#files[@]}
9
10 # get a random array element
11 wallpaper="${files[RANDOM % n]}"
12
13 # set it as wallpaper
14 qiv -z $wallpaper

```



Change wallpaper every 5 mins?

- <https://www.tutorialspoint.com/unix/unix-using-arrays.htm>

Part III

Linux Programming Environment

3 C Programming Environment

Program Languages

Machine code

The *binary numbers* that the CPUs can understand.

100111000011101111001111 ... and so on ...

People don't think in numbers.

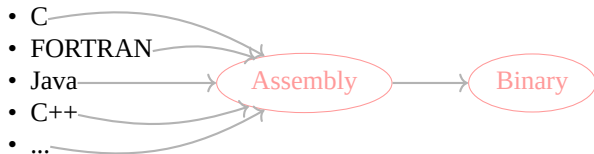
Assembly language — friendly to humans

```
1  MOV A,47 ;1010 1111
2  ADD A,B  ;0011 0111
3  HALT     ;0111 0110
```

Assemblers translate the ASM programs to machine code

High level languages

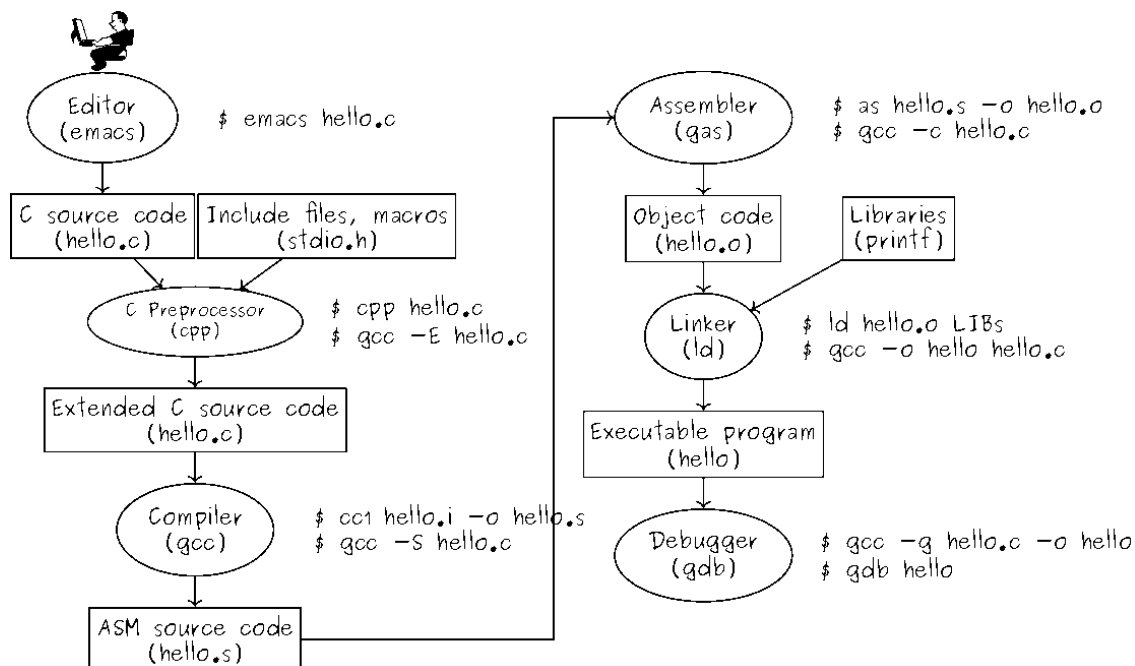
Even easier to understand by humans. Examples:



Compilers do the translation work

3.1 The Tool Chain

Compilation



Source code written by programmer in high-level language, in our case in C. We write c source code with a *text editor*, such as emacs, vim, etc.

Preprocessing is the first pass of any C compilation. It processes include-files, conditional compilation instructions and macros.

cpp The GNU C preprocessor

```
$ gcc -E hello.c
```

Compilation is the second pass. It takes the output of the preprocessor, and the source code, and generates assembly source code.

gcc/g++ GNU C/C++ compiler

```
$ gcc -S hello.c
```

Assembly is the third stage of compilation. It takes the assembly source code and produces an assembly listing with offsets. The assembler output is stored in an object file.

as the portable GNU assembler

```
$ gcc -c hello.c
```

Linking is the final stage of compilation. It combines object code with predefined routines from libraries and produces the executable program.

ld The GNU linker

```
$ gcc hello.c -lm
```

Wrapper The whole compilation process is usually not done ‘by hand’, but using a wrapper program that combines the functions of preprocessor(cpp), compiler(gcc/g++), assembler(as) and linker(ld).

```
$ gcc -Wall hello.c -lm -o hello
```

Compiler vs. Interpreter

hello.c

```
1 #include <stdio.h>
2 int main()
3 {
4     printf("Hello, world!\n");
5     return 0;
6 }
```

```
$ gcc -o hello hello.c
```

```
$ ./hello
```

hello.sh

```
1 #!/bin/bash
2 echo 'Hello, world!'
```

```
$ chmod +x hello.sh
```

```
$ ./hello.sh
```

hello.py

```
1 #!/usr/bin/python
2 print "Hello, world!"
```

```
$ chmod +x hello.py
```

```
$ ./hello.py
```

3.2 Header Files

Header Files

Why?

```
1 #include "add.h"
2
3 int triple(int x)
4 {
5     return add(x, add(x,x));
6 }
```

- Ensure everyone use the same code
- Easy to share, upgrade, reuse

In the header files...

Why not?

```
1 int add(int, int);
2
3 int triple(int x)
4 {
5     return add(x, add(x, x));
6 }
```

- function declarations
- macro definitions
- constants
- system wide global variables

```
$ ls /usr/include/
```

3.3 Library Files

Library Files

Static libraries .a files. Very old ones, but still alive.

```
$ find /usr/lib -name "*.a"
```

Shared libraries .so files. The preferred ones.

```
$ find /usr/lib -name "*.so.*"
```

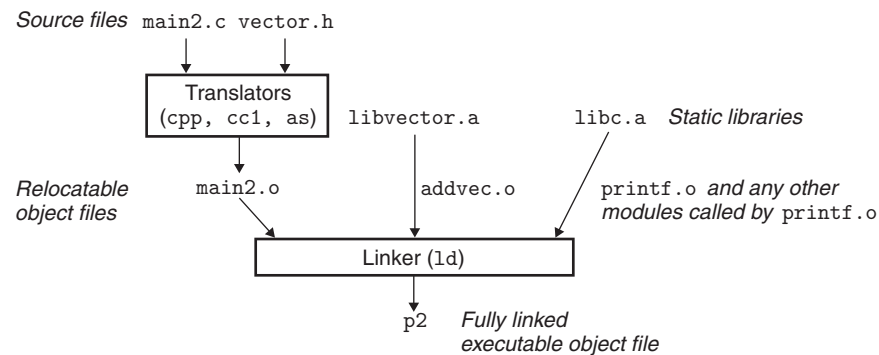
Examples:

```
$ gcc -o hello hello.c /usr/lib/libm.a
```

```
$ gcc -o hello hello.c -lm
```

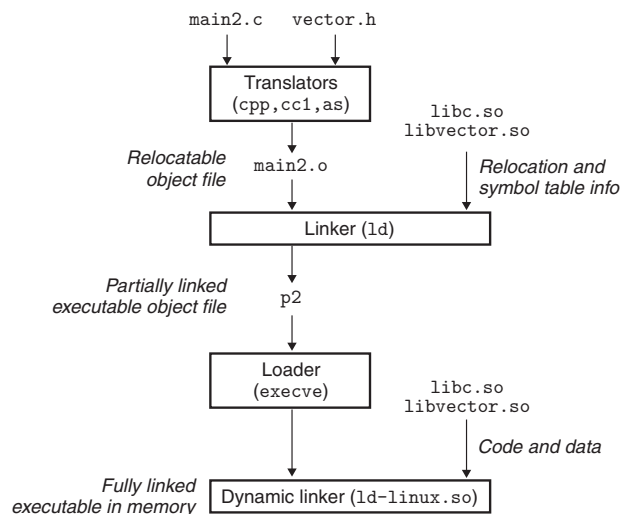
Static Linking

- The entire program and all data of a process must be in physical memory for the process to execute
- The size of a process is thus limited to the size of physical memory



Dynamic Linking

- Only one copy in memory
- Don't have to re-link after a library update



Build A Static Library

Source codes

main.c

```
1 | #include "lib.h"
2 |
3 | int main(int argc, char* argv[])
4 | {
5 |     int i=1;
6 |
7 |     for (; i<argc; i++)
8 |     {
9 |         hello(argv[i]);
10 |        hi(argv[i]);
11 |    }
12 |    return 0;
13 | }
```

lib.h

```
1 | #include <stdio.h>
2 |
3 | void hello(char *);
4 | void hi(char *);
```

hello.c

```
1 | #include <stdio.h>
2 |
3 | void hello(char *arg)
4 | {
5 |     printf("Hello, %s!\n", arg);
6 | }
```

hi.c

```
1 | #include <stdio.h>
2 |
3 | void hi(char *arg)
4 | {
5 |     printf("Hi, %s!\n", arg);
6 | }
```

Build A Static Library

Step by step

1. Get *hello.o* and *hi.o*
\$ gcc -c hello.c hi.c
2. Put *.o into *libhi.a*
\$ ar crv libhi.a hello.o hi.o
3. Use *libhi.a*
\$ gcc main.c libhi.a

Build A Static Library

Makefile

```
1 | main: main.c lib.h libhi.a
2 |     gcc -Wall -o main main.c libhi.a
3 |
4 | libhi.a: hello.o hi.o
5 |     ar crv libhi.a hello.o hi.o
6 |
7 | hello.o: hello.c
8 |     gcc -Wall -c hello.c
9 |
10 | hi.o: hi.c
11 |     gcc -Wall -c hi.c
12 |
13 | clean:
14 |     rm -f *.o *.a main
```

Build A Shared Library

Source codes

hello.c

```

1 | #include "hello.h"
2 |
3 | int main(int argc, char *argv[])
4 | {
5 |     if (argc != 2)
6 |         printf ("Usage: %s needs an argument.\n", argv[0]);
7 |     else
8 |         hi(argv[1]);
9 |     return 0;
10 | }

```

hi.c

hello.h

```

1 | #include <stdio.h>
2 | #include <stdlib.h>
3 |
4 | int hi(char*);

```

```

1 | #include "hello.h"
2 |
3 | int hi(char* s)
4 | {
5 |     printf ("Hi, %s\n",s);
6 |     return 0;
7 | }

```

Build A Shared Library

Step by step

1. Get *hi.o*
\$ gcc -fPIC -c hi.c
2. Get *libhi.so*
\$ gcc -shared -o libhi.so hi.o
3. Use *libhi.so*
\$ gcc -L. -Wl,-rpath=. hello.c -lhi
4. Check it
\$ ldd a.out

Build A Shared Library

Makefile

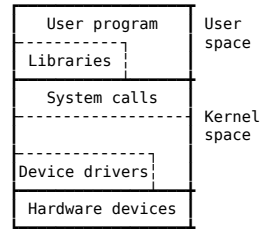
```

1 | # http://www.cprogramming.com/tutorial/shared-libraries-linux-gcc.html
2 | # http://tldp.org/HOWTO/Program-Library-HOWTO/shared-libraries.html
3 | #
4 | # gcc -fPIC -c hi.c
5 | # gcc -shared -o libhi.so hi.o
6 | # gcc -L/current/dir -Wl,option -Wall -o hello hello.c -lhi
7 | #
8 | # -L          - tells ld where to search libraries
9 | # -Wl,option - pass option as an option to the linker (ld)
10 | # -rpath=dir - Add a directory to the runtime library search path
11 |
12 | hello: hello.c hello.h libhi.so
13 |     gcc -L. -Wl,-rpath=. -Wall -o hello hello.c -lhi
14 | libhi.so: hi.o hello.h
15 |     gcc -shared -o libhi.so hi.o
16 | hi.o: hi.c hello.h
17 |     gcc -fPIC -c hi.c
18 | clean:
19 |     rm *.o *.so hello

```

GNU C Library

```
Linux API > POSIX API
$ man 7 libc
$ man 3 intro
$ man gcc
$ info gcc
🕒 sudo apt install gcc-doc
```



3.4 Error Handling

errno.h

```
1 #include <stdio.h>
2 #include <stdlib.h>
3 #include <sys/stat.h>
4 #include <fcntl.h>
5
6 int main(int argc, char *argv[])
7 {
8     if( open(argv[1], O_RDONLY) == -1 ){
9         perror("open");
10        exit(EXIT_FAILURE);
11    }
12    return 0;
13 }
14
15 /* Local Variables: */
16 /* compile-command: "gcc -Wall -Wextra perror.c -o perror" */
17 /* End: */
```

```
$ man errno
$ man errno.h
$ man perror
• [Advanced programming in the UNIX environment, Sec. 1.7]
• https://stackoverflow.com/questions/30078281/raise-error-in-a-bash-script
```

3.5 The Make Utility

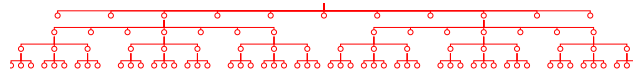
The Make Utility

To compile a single C program:

```
$ gcc hello.c -o hello
```

✔ OK. But...

What if you have a large project with 1000+ files?



Linux 4.9 source tree: 3799 directories, 55877 files

make: help you maintain your programs.

Makefile

```
1 target: dependencies
2 |———TAB——> command
```

Example

```
1 hello: hello.c
2 |———TAB——> gcc -o hello hello.c
```

```
$ info make makefiles
```

Makefile

```
1 edit: main.o kbd.o command.o display.o \  
2         insert.o search.o files.o utils.o  
3         gcc -Wall -o edit main.o kbd.o command.o display.o \  
4         insert.o search.o files.o utils.o  
5  
6 main.o: main.c defs.h  
7         gcc -c -Wall main.c  
8  
9 kbd.o : kbd.c defs.h command.h  
10        gcc -c -Wall kbd.c  
11  
12 command.o: command.c defs.h command.h  
13        gcc -c -Wall command.c  
14  
15 display.o : display.c defs.h buffer.h  
16        gcc -c -Wall display.c  
17  
18 insert.o: insert.c defs.h buffer.h  
19        gcc -c -Wall insert.c  
20  
21 search.o: search.c defs.h buffer.h  
22        gcc -c -Wall search.c  
23  
24 files.o: files.c defs.h buffer.h command.h  
25        gcc -c -Wall files.c  
26  
27 utils.o: utils.c defs.h  
28        gcc -c -Wall utils.c  
29  
30  
31 clean:  
32        rm edit main.o kbd.o command.o display.o \  
33        insert.o search.o files.o utils.o
```

```
./  
├── command.c  
├── display.c  
├── files.c  
├── insert.c  
├── kbd.c  
├── main.c  
├── search.c  
├── utils.c  
├── buffer.h  
├── command.h  
├── defs.h  
└── Makefile
```

3.6 Version Control

git

To create a new local git repo

In your source code directory, do:

```
$ git init  
$ git add .  
$ git commit -m "something to say..."
```

To clone a remote repo

Example:

```
$ git clone https://github.com/wx672/lecture-notes.git  
$ git clone https://github.com/wx672/dotfile.git
```

Most commonly used git Commands

```
$ git add filename[s]  
$ git rm filename[s]  
$ git commit  
$ git status      $ git log      $ git diff  
$ git push        $ git pull  
$ git help {add,rm,commit,...}
```

```
$ man gittutorial
$ man gittutorial-2
```

```
sudo apt install git
https://github.com
```

3.7 Manual Pages

Man page

Layout

```
1  NAME
2      A one-line description of the command.
3  SYNOPSIS
4      A formal description of how to run it and what
5      command line options it takes.
6  DESCRIPTION
7      A description of the functioning of the command.
8  EXAMPLES
9      Some examples of common usage.
10 SEE ALSO
11     A list of related commands or functions.
12 BUGS
13     List known bugs.
14 AUTHOR
15     Specify your contact information.
16 COPYRIGHT
17     Specify your copyright information.
```

Man Page

Groff source code

```
1 .\" Text automatically generated by txt2man
2  .TH untitled "06 August 2019" "" ""
3  .SH NAME
4  \fBA one-line description of the command.
5  .SH SYNOPSIS
6  .nf
7  .fam C
8  \fBA formal description of how to run it and what command line options it takes.
9  .fam T
10 .fi
11 .fam T
12 .fi
13 .SH DESCRIPTION
14 \fBA description of the functioning of the command.
15 .SH EXAMPLES
16 Some examples of common usage.
17 .SH SEE ALSO
18 \fBA list of related commands or functions.
19 .SH BUGS
20 List known bugs.
21 .SH AUTHOR
22 Specify your contact information.
23 .SH COPYRIGHT
24 Specify your copyright information.
```

```
$ man 7 groff
$ man txt2man
$ man a2x
$ ls /usr/share/man
```

3.8 A Sample GNU Package

How to “Do one thing, and do it well”?

```
$ apt source hello
```

4 The Linux Environment

Command Line Options

getopt.c

```
1 | #include <stdio.h>
2 | #include <unistd.h>
```

```

3
4 int main(int argc, char* argv[]) {
5     int opt;
6
7     while ((opt = getopt(argc, argv, "hf:l")) != -1) {
8         switch (opt) {
9             case 'h':
10                printf("Usage: %s [-h] [-f file] [-l]\n", argv[0]);
11                break;
12             case 'l':
13                printf("option: %c\n", opt);
14                break;
15             case 'f':
16                printf("filename: %s\n", optarg);
17                break;
18            }
19        }
20        return 0;
21    }
22
23    /* Local Variables: */
24    /* compile-command: "gcc -Wall -Wextra getopt.c -o getopt" */
25    /* End: */

```

```
$ man 3 getopt
```

Command Line Options

getopt.sh

```

1 #!/bin/bash
2
3 while getopts hf:l OPT; do
4     case $OPT in
5         h) echo "usage: `basename $0` [-h] [-f file] [-l]"
6             exit 1 ;;
7         l) echo "option: l" ;;
8         f) echo "filename: $OPTARG" ;;
9     esac
10 done

```

```

$ ./getopt.sh -h
$ ./getopt.sh -lf filename
$ ./getopt.sh -l -f filename
$ ./getopt.sh -f filename -l

```

Environment Variable

```

1  #include <stdlib.h>
2  #include <stdio.h>
3
4  extern char** environ;
5
6  int main() {
7      char** env = environ;
8
9      while (*env) {
10         printf("%s\n", *env);
11         env++;
12     }
13
14     return 0;
15 }
16
17 /* Local Variables: */
18 /* compile-command: "gcc -Wall -Wextra env.c
19 ↪ -o env" */
19 /* End: */

```

```

$ env
$ man 3 getenv
$ man 3 putenv

```

[Beginning linux programming, p.147, Sec. The environ Variable]

Time and Date

```

1  #include <time.h>
2  #include <stdio.h>
3
4  int main(void)
5  {
6      time_t t = time(NULL); /* long int */
7
8      printf("epoch time:\t%ld\n",t);
9      printf("calendar time:\t%s", ctime(&t));
10
11     return 0;
12 }
13
14 /* Local Variables: */
15 /* compile-command: "gcc -Wall -Wextra time.c -o time" */
16 /* End: */

```

- January 1 1970 — start of the Unix epoch

```
$ man 3 time
```

```
$ man 3 ctime
```

Temporary Files

mkstemp.c

```
1 #define _GNU_SOURCE
2 #include <stdlib.h>
3 #include <unistd.h>
4 #include <stdio.h>
5
6 int main(int argc, char *argv[])
7 {
8     char c, *f;
9
10    asprintf(&f, "%sXXXXXX", argv[1]);
11    int tmp = mkstemp(f);
12
13    while ( read(0, &c, 1) == 1)
14        write(tmp, &c, 1);
15
16    unlink(f);
17    free(f);
18    return 0;
19 }
20
21 /* Local Variables: */
22 /* compile-command: "gcc -Wall -Wextra mkstemp.c -o
23    ↪ mkstemp" */
24 /* End: */
```

mktemp.sh

```
1 #!/bin/bash
2
3 tmp=$(mktemp)
4
5 while read LINE; do
6     echo $LINE >> $tmp
7 done
8
9 $ man 3 mkstemp
10 $ man 3 tmpfile
11 $ man 3 asprintf
```

Logging

syslog.c

```
1 #include <syslog.h>
2 #include <sys/stat.h>
3 #include <fcntl.h>
4
5 int main(int argc, char *argv[])
6 {
7     if( open(argv[1], O_RDONLY) < 0 )
8         syslog(LOG_ERR | LOG_USER, "%s - %m\n", argv[1]);
9     else
10         syslog(LOG_INFO | LOG_USER, "%s - %m\n", argv[1]);
11     return 0;
12 }
13
14 /* Local Variables: */
15 /* compile-command: "gcc -Wall -Wextra syslog.c -o syslog" */
16 /* End: */
```

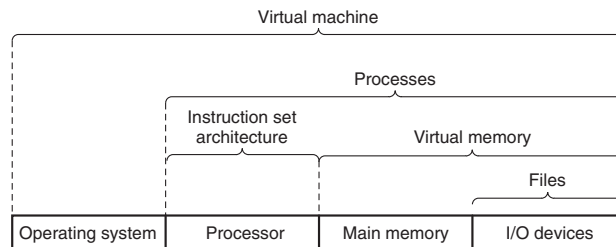
logger.sh

```
1 #!/bin/bash
2
3 [[ -f "$1" ]] && logger "$1 exists." || logger "$1 not found."
```

5 OS Basics

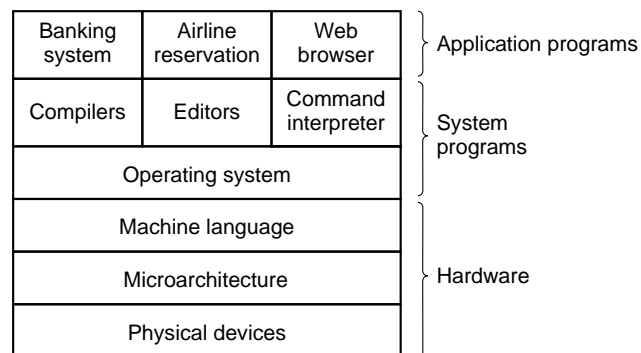
Abstractions

To hide the complexity of the actual implementations



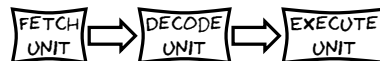
See also: [Computer Systems: A Programmer's Perspective, Sec. 1.9.2, The Importance of Abstractions in Computer Systems].

A Computer System



5.1 Hardware

CPU Working Cycle



1. Fetch the first instruction from memory
2. Decode it to determine its type and operands
3. execute it

Special CPU Registers

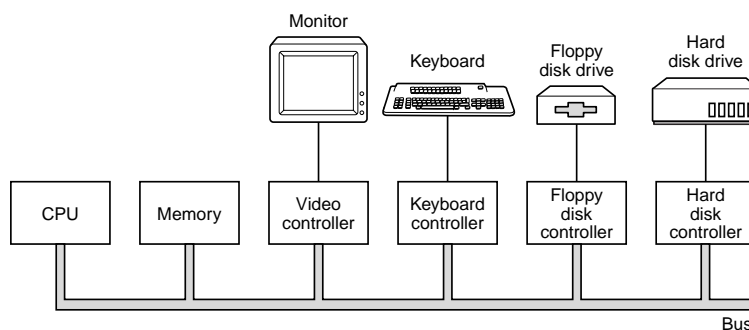
Program counter (PC): keeps the memory address of the next instruction to be fetched

Stack pointer (SP): points to the top of the current stack in memory

Program status (PS): holds

- condition code bits
- processor state

System Bus



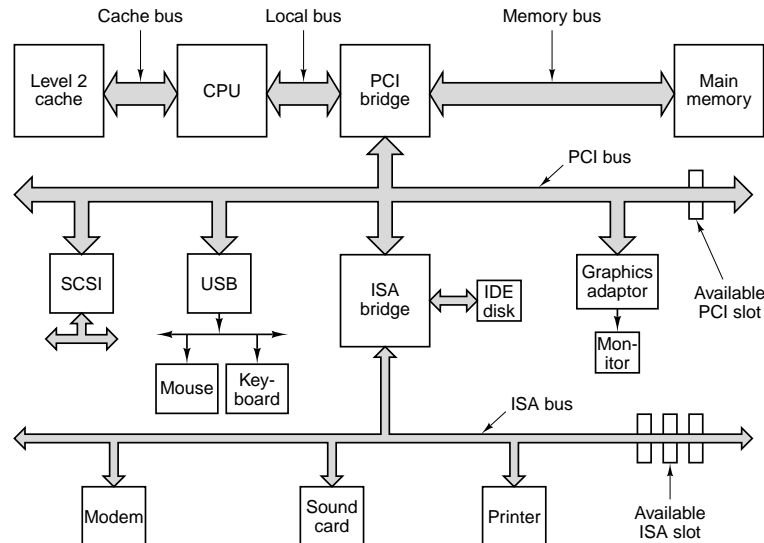
Address Bus: specifies the memory locations (addresses) for the data transfers

Data Bus: holds the data transferred. Bidirectional

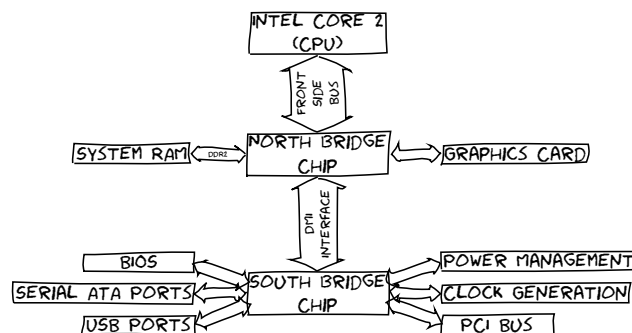
Control Bus: contains various lines used to route timing and control signals throughout the system

Controllers and Peripherals

- Peripherals are real devices controlled by controller chips
- Controllers are processors like the CPU itself, have control registers
- Device driver writes to the registers, thus control it
- Controllers are connected to the CPU and to each other by a variety of buses



Motherboard Chipsets



See also: *Motherboard Chipsets And The Memory Map*¹.

- The CPU doesn't know what it's connected to
 - CPU test bench? network router? toaster? brain implant?
- The CPU talks to the outside world through its pins
 - some pins to transmit the physical memory address
 - other pins to transmit the values
- The CPU's gateway to the world is the *front-side bus*

Intel Core 2 QX6600

- 33 pins to transmit the physical memory address
 - so there are 2^{33} choices of memory locations
- 64 pins to send or receive data
 - so data path is 64-bit wide, or 8-byte chunks

¹<http://duartes.org/gustavo/blog/post/motherboard-chipsets-memory-map>

This allows the CPU to physically address 64GB of memory ($2^{33} \times 8B$)

See also: *Datasheet for Intel Core 2 Quad-Core Q6000 Sequence* ².

Some physical memory addresses are mapped away!

- only the addresses, not the spaces
- Memory holes
 - 640 KiB ~ 1 MiB
 - /proc/iomem

Memory-mapped I/O

- BIOS ROM
- video cards
- PCI cards
- ...

This is why 32-bit OSes have problems using 4 GiB of RAM.

0xFFFFFFFF	Reset vector	4GB
0xFFFFFFF0	JUMP to 0xF0000	4GB-16B
	Unaddressable memory, real mode is limited to 1MB.	
0x100000	System BIOS	1MB
0xF0000	Ext. System BIOS	960KB
0xE0000	Expansion Area (maps ROMs for old peripheral cards)	896KB
0xC0000	Legacy Video Card Memory Access	768KB
0xA0000	Accessible RAM (640KB is enough for anyone – old DOS area)	640KB
0		0

the northbridge

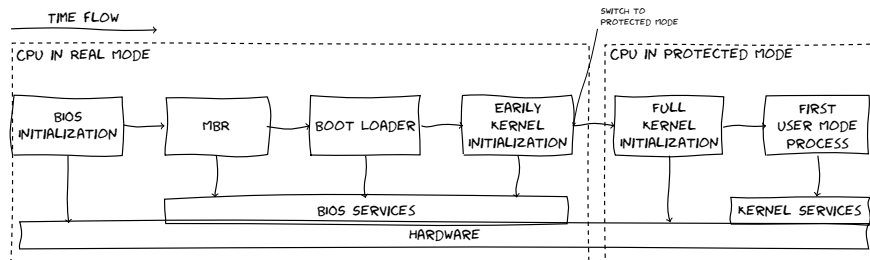
1. receives a physical memory request
 2. decides where to route it
 - to RAM? to video card? to ...?
 - decision made via the *memory address map*
- When is the memory address map built? `setup()`.

5.2 Bootstrapping

Bootstrapping

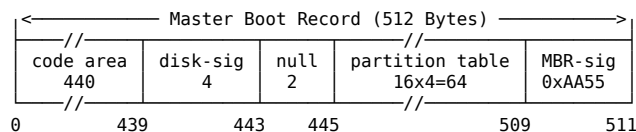
Can you pull yourself up by your own bootstraps?

A computer cannot run without first loading software but must be running before any software can be loaded.



Intel x86 Bootstrapping

1. BIOS (0xfffffff0)
 - ➡ POST ➡ HW init ➡ Find a boot device (FD,CD,HD...) ➡ Copy sector zero (MBR) to RAM (0x00007c00)
2. MBR – the first 512 bytes, contains
 - Small code (< 446 B), e.g. GRUB stage 1, for loading GRUB stage 2
 - the primary partition table ($16 \times 4 = 64 B$)
 - its job is to load the second-stage boot loader.
3. GRUB stage 2 — load the OS kernel into RAM
4. startup
5. init — the first user-space process



\$ `sudo hd -n512 /dev/sda`

- https://en.wikipedia.org/wiki/Unified_Extensible_Firmware_Interface

²<http://download.intel.com/design/processor/datashts/31559205.pdf>

5.3 Interrupt

Why Interrupt?

While a process is reading a disk file, can we do...

```
1 while(!done_reading_a_file())
2 {
3     let_CPU_wait();
4     // or...
5     lend_CPU_to_others();
6 }
7 operate_on_the_file();
```

Modern OS are Interrupt Driven

HW INT by sending a signal to CPU

SW INT by executing a *system call*

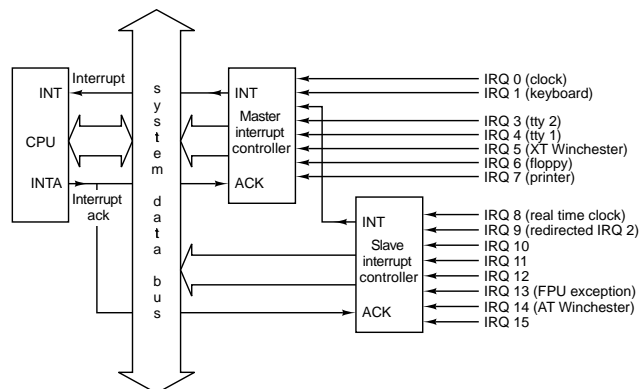
Trap (exception) is a software-generated INT caused by an error or by a specific request from an user program

Interrupt vector is an array of pointers to the memory addresses of *interrupt handlers*. This array is indexed by a unique device number

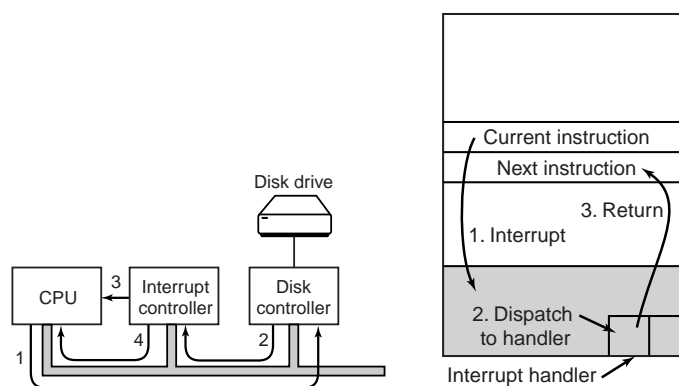
```
$ less /proc/devices
```

```
$ less /proc/interrupts
```

Programmable Interrupt Controllers

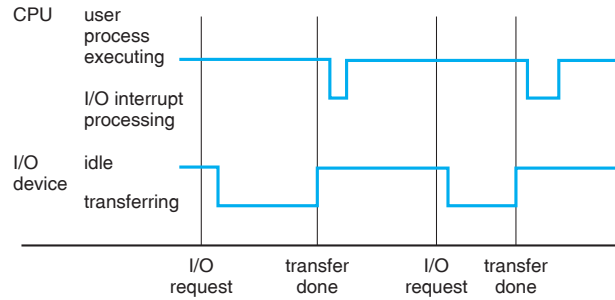


Interrupt Processing



- [Modern Operating Systems, Sec. 1.3.5, I/O Devices]

Interrupt Timeline



5.4 System Calls

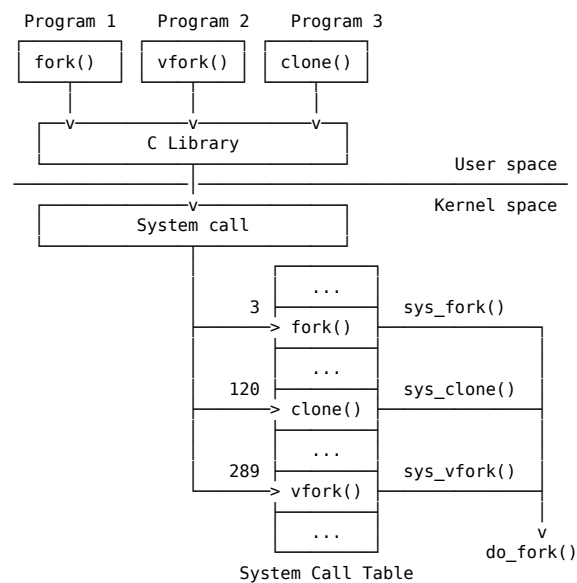
System Calls

A System Call

- is how a program requests a service from an OS kernel
- provides the interface between a process and the OS

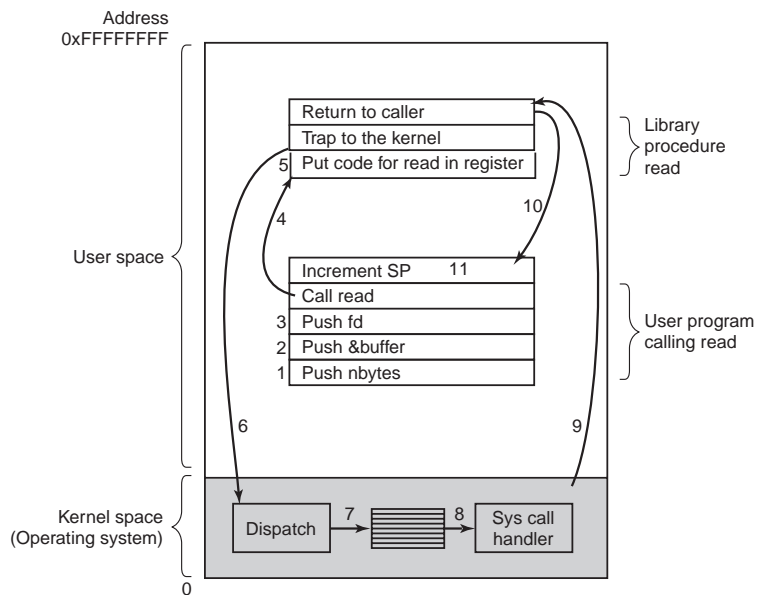
\$ man 2 intro

\$ man 2 syscalls



The 11 steps in making a system call

read(fd, buffer, nbytes)



Example

Linux `INT 80h`

Interrupt Vector Table: The very first 1KiB of x86 memory.

- 256 entries \times 4B = 1KiB
- Each entry is a complete memory address (segment:offset)
- It's populated by Linux and BIOS
- Slot 80h: address of the kernel services dispatcher (☛ sys-call table)

Example

```

1  Msg: db 'Hello, world'
2  MsgLen: equ $-Msg
3  mov eax,4      ; sys_write syscall = 4
4  mov ebx,1      ; 1 = STDOUT
5  mov ecx,Msg    ; offset of the message
6  mov edx,MsgLen ; length of string
7  int 80h        ; call the kernel

```

```

$ nasm -f elf64 hello.asm -o hello.o
$ ld hello.o -o hello
$ ./hello

```

Process management	
Call	Description
<code>pid = fork()</code>	Create a child process identical to the parent
<code>pid = waitpid(pid, &statloc, options)</code>	Wait for a child to terminate
<code>s = execve(name, argv, environp)</code>	Replace a process' core image
<code>exit(status)</code>	Terminate process execution and return status

File management	
Call	Description
<code>fd = open(file, how, ...)</code>	Open a file for reading, writing or both
<code>s = close(fd)</code>	Close an open file
<code>n = read(fd, buffer, nbytes)</code>	Read data from a file into a buffer
<code>n = write(fd, buffer, nbytes)</code>	Write data from a buffer into a file
<code>position = lseek(fd, offset, whence)</code>	Move the file pointer
<code>s = stat(name, &buf)</code>	Get a file's status information

Directory and file system management	
Call	Description
<code>s = mkdir(name, mode)</code>	Create a new directory
<code>s = rmdir(name)</code>	Remove an empty directory
<code>s = link(name1, name2)</code>	Create a new entry, name2, pointing to name1
<code>s = unlink(name)</code>	Remove a directory entry
<code>s = mount(special, name, flag)</code>	Mount a file system
<code>s = umount(special)</code>	Unmount a file system

Miscellaneous	
Call	Description
<code>s = chdir(dirname)</code>	Change the working directory
<code>s = chmod(name, mode)</code>	Change a file's protection bits
<code>s = kill(pid, signal)</code>	Send a signal to a process
<code>seconds = time(&seconds)</code>	Get the elapsed time since Jan. 1, 1970

Fig. 1-18. Some of the major POSIX system calls. The return code *s* is -1 if an error has occurred. The return codes are as follows: *pid* is a process id, *fd* is a file descriptor, *n* is a byte count, *position* is an offset within the file, and *seconds* is the elapsed time. The parameters are explained in the text.

System Call Examples

`fork()`

```

1  #include <stdio.h>
2  #include <unistd.h>
3
4  int main ()
5  {
6      printf("Hello World!\n");
7      fork();
8      printf("Goodbye Cruel World!\n");
9      return 0;
10 }
```

`$ man 2 fork`

`execve()`

```

1 #include <stdio.h>
2 #include <unistd.h>
3
4 int main()
5 {
6     printf("Hello World!\n");
7     if( fork() != 0 )
8         printf("I am the parent process.\n");
9     else {
10         printf("A child is listing the directory contents...\n");
11         execl("/bin/ls", "ls", "-al", NULL);
12     }
13     return 0;
14 }

```

\$ man 2 execve

\$ man 3 exec

Quoted from [stackoverflow](#): What is the difference between the functions of the *exec* family of system calls:

There is no *exec* system call — this is usually used to refer to all the *execXX* calls as a group. They all do essentially the same thing: loading a new program into the current process, and provide it with arguments and environment variables. The differences are in how the program is found, how the arguments are specified, and where the environment comes from.

- The calls with *v* in the name take an array parameter to specify the `argv[]` array (*vector*) of the new program.
- The calls with *l* in the name take the arguments of the new program as a variable-length argument *list* to the function itself.
- The calls with *e* in the name take an extra argument to provide the *environment* of the new program; otherwise, the program inherits the current process's environment.
- The calls with *p* in the name search the *PATH* environment variable to find the program if it doesn't have a directory in it (i.e. it doesn't contain a `/` character). Otherwise, the program name is always treated as a path to the executable.

Part IV

Working With Files

6 Files

File

A logical view of information storage

User's view

A file is the smallest storage unit on disk.

- Data cannot be written to disk unless they are within a file

UNIX view

Each file is a sequence of 8-bit bytes

- It's up to the application program to interpret this byte stream.

File

What is stored in a file?

Source code, object files, executable files, shell scripts, PostScript...

Different type of files have different structure

- UNIX looks at contents to determine type
 - Shell scripts** start with “#!”
 - PDF** start with “%PDF . . .”
 - Executables** start with *magic number*
- Windows uses file naming conventions
 - executables** end with “.exe” and “.com”
 - MS-Word** end with “.doc”
 - MS-Excel** end with “.xls”

File Types

Regular files: ASCII, binary

Directories: Maintaining the structure of the FS

In UNIX, everything is a file

Character special files: I/O related, such as terminals, printers ...

Block special files: Devices that can contain file systems, i.e. disks

Disks — logically, linear collections of blocks; disk driver translates them into physical block addresses

File Operations

POSIX file system calls

creat(name, mode)	read(fd, buffer, byte_count)
open(name, flags)	write(fd, buffer, byte_count)
close(fd)	lseek(fd, offset, whence)
link(oldname, newname)	chown(name, owner, group)
unlink(name)	fchown(fd, owner, group)
truncate(name, size)	chmod(name, mode)
ftruncate(fd, size)	fchmod(fd, mode)
stat(name, buffer)	utimes(name, times)
fstat(fd, buffer)	

write()

```

1 #include <unistd.h>
2
3 int main(void)
4 {
5     write(1, "Hello, world!\n", 14);
6
7     return 0;
8 }
9
10 /* Local Variables: */
11 /* compile-command: "gcc -Wall -Wextra write.c -o
   ↪ write" */
12 /* End: */

$ man 2 write
$ man 3 write

```

read()

```

1 #include <unistd.h>
2
3 int main(void)
4 {
5     char buffer[10];
6
7     read(0, buffer, 10);
8
9     write(1, buffer, 10);
10
11     return 0;
12 }
13
14 /* Local Variables: */
15 /* compile-command: "gcc -Wall
   ↪ -Wextra read.c -o read" */
16 /* End: */

$ man 2 read
$ man 3 read

```

- No need to open() STDIN, STDOUT, and STDERR

cp

```

1 #include <sys/types.h> /* include necessary header files */
2 #include <fcntl.h>
3 #include <stdlib.h>
4 #include <unistd.h>
5
6 #define BUF_SIZE 4096
7 #define OUTPUT_MODE 0700
8
9 int main(int argc, char *argv[])
10 {
11     int in, out, rbytes, wbytes;
12     char buf[BUF_SIZE];
13
14     if (argc != 3) exit(1);
15
16     if ( (in = open(argv[1], O_RDONLY)) < 0 ) exit(2);
17
18     if ( (out = creat(argv[2], OUTPUT_MODE)) < 0 ) exit(3);
19
20     while (1) { /* Copy loop */
21         if ( (rbytes = read(in, buf, BUF_SIZE)) <= 0 ) break;
22         if ( (wbytes = write(out, buf, rbytes)) <= 0 ) exit(4);
23     }
24 }

```

```

25     close(in); close(out);
26     if (rbytes == 0) exit(0); /* no error on last read */
27     else exit(5);             /* error on last read */
28 }
29
30 /* Local Variables: */
31 /* compile-command: "gcc -Wall -Wextra cp-syscall.c -o cp-syscall" */
32 /* End: */

```

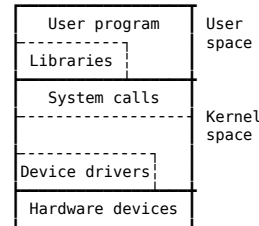
stdio — The Standard I/O Library

System calls: `open()`, `read()`, `write()`, `close()`...

Library functions: `fopen()`, `fread()`, `fwrite()`, `fclose()`...

Avoid calling syscalls directly as much as you can

- Portability
- Buffered I/O



`open()` vs. `fopen()`

`open()`

```

1  #include <unistd.h>
2  #include <sys/stat.h>
3  #include <fcntl.h>
4  #include <stdio.h>
5
6  int main()
7  {
8      char c;
9      int in;
10     in = open("/tmp/1m.test", O_RDONLY);
11
12     while (read(in, &c, 1) == 1);
13
14     return 0;
15 }
16
17 /* Local Variables: */
18 /* compile-command: "gcc -Wall -Wextra
19 ↪ open.c -o open" */
20 /* End: */

```

```
$ strace -c ./open
```

```
$ dd if=/dev/zero of=/tmp/1m.test bs=1k count=1024
```

- <https://stackoverflow.com/questions/1658476/c-fopen-vs-open>

`fopen()` — Buffered I/O

```

1  #include <stdio.h>
2
3  int main(void)
4  {
5      FILE *stream;
6
7      stream = fopen("/tmp/1m.test", "r");
8
9      while ( fgetc(stream) != EOF );
10
11     fclose(stream);
12
13     return 0;
14 }
15
16 /* Local Variables: */
17 /* compile-command: "gcc -Wall -Wextra
18 ↪ fopen.c -o fopen" */
19 /* End: */

```

```
$ strace -c ./fopen
```

cp — With stdio

```

1  #include <stdio.h>
2  #include <stdlib.h>
3
4  int main(int argc, char *argv[])

```

```

5 {
6     FILE *in, *out;
7     int c=0;
8
9     if (argc != 3) exit(1);
10
11     in = fopen(argv[1], "r");
12     out = fopen(argv[2], "w");
13
14     while( (c = fgetc(in)) != EOF )
15         fputc(c, out);
16
17     return 0;
18 }
19
20 /* Local Variables: */
21 /* compile-command: "gcc -Wall -Wextra cp-libc.c -o cp-libc" */
22 /* End: */

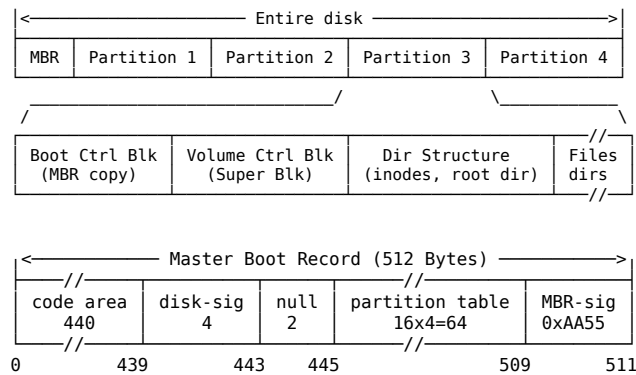
```

□ Try fread()/fwrite() instead.

- <https://stackoverflow.com/questions/32742430/is-getc-a-macro-or-a-function>
- <https://stackoverflow.com/questions/9104568/macro-vs-function-in-c>

File System Implementation

A typical file system layout



On-Disk Information Structure

Boot block a MBR copy

Superblock Contains volume details

number of blocks	size of blocks
free-block count	free-block pointers
free FCB count	free FCB pointers

I-node Organizes the files *FCB (File Control Block)*, contains file details (metadata).

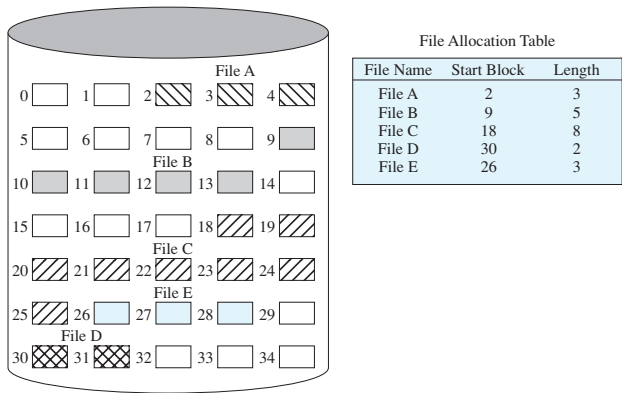
Superblock

Keeps information about the file system

- Type — ext2, ext3, ext4...
- Size
- Status — how it's mounted, free blocks, free inodes, ...
- Information about other metadata structures

```
$ sudo dumpe2fs /dev/sda1 | less
```

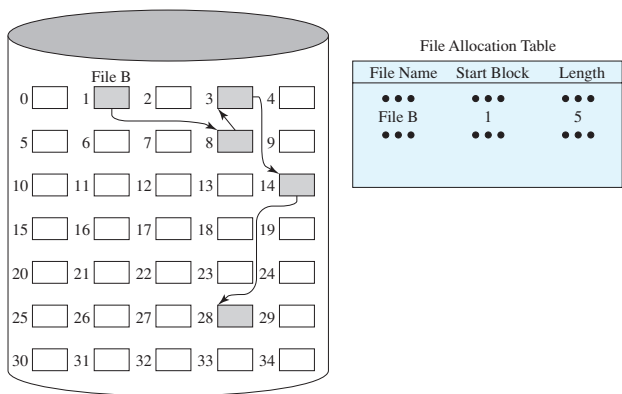
Implementing Files



Contiguous Allocation

- 😊 simple
- 😊 good for read only

☹ fragmentation



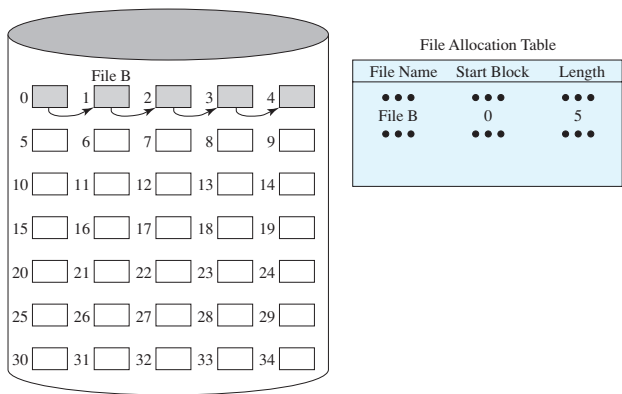
Linked List (Chained) Allocation

A pointer in each disk block

- 😊 no waste block
- ☹ slow random access

☹ not 2^n

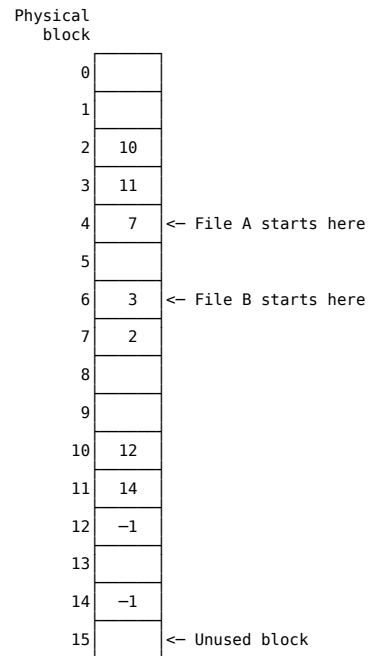
Linked List (Chained) Allocation Though there is no external fragmentation, consolidation is still preferred.



FAT: Linked list allocation with a table in RAM

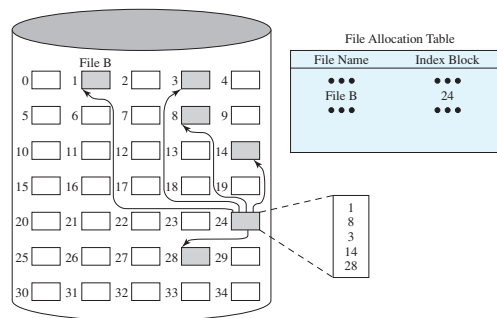
- Taking the pointer out of each disk block, and putting it into a table in memory
- fast random access (chain is in RAM)
- is 2^n
- the entire table must be in RAM

$$disk \nearrow \Rightarrow FAT \nearrow \Rightarrow RAM_{used} \nearrow$$



See also: [File Allocation Table — Wikipedia, The Free Encyclopedia, Wikipedia:FAT].

Indexed Allocation

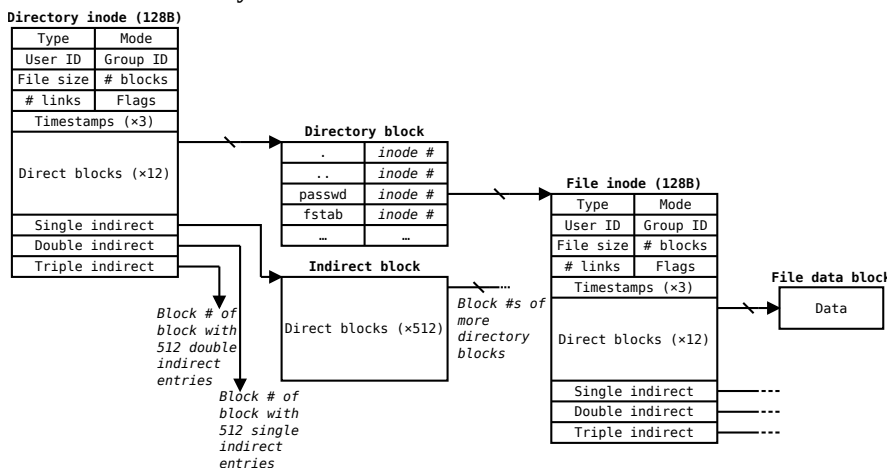


I-node A data structure for each file. An i-node is in memory *only if* the file is open

$$files_{opened} \nearrow \Rightarrow RAM_{used} \nearrow$$

See also: [Inode — Wikipedia, The Free Encyclopedia, Wikipedia:inode].

UNIX Treats a Directory as a File

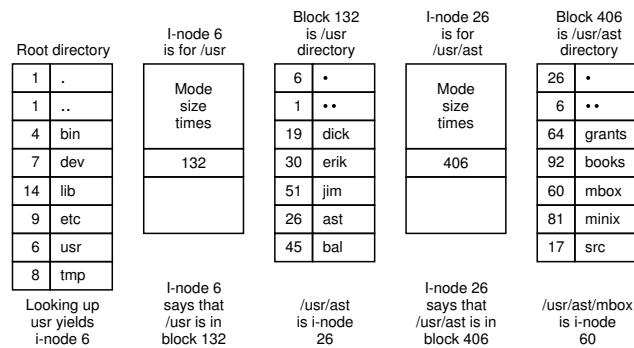


`open()`

Why? To avoid constant searching

- Without `open()`, every file operation involves searching the directory for the file.

The steps in looking up `/usr/ast/mbox`



`fd open(pathname, flags)`

A per-process *open-file table* is kept in the OS

- upon a successful `open()` syscall, a new entry is added into this table
- indexed by *file descriptor (fd)*
- `close()` to remove an entry from the table

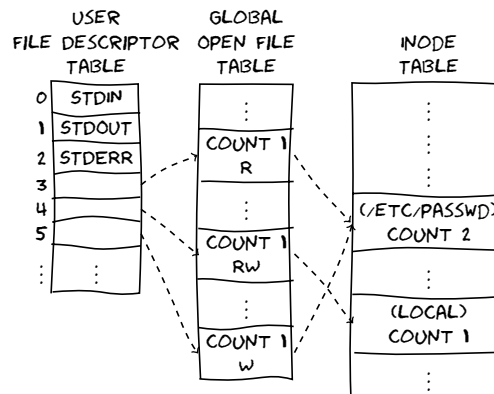
To see files opened by a process, e.g. `init`

`$ lsof -p 1`

`$ man 2 open`

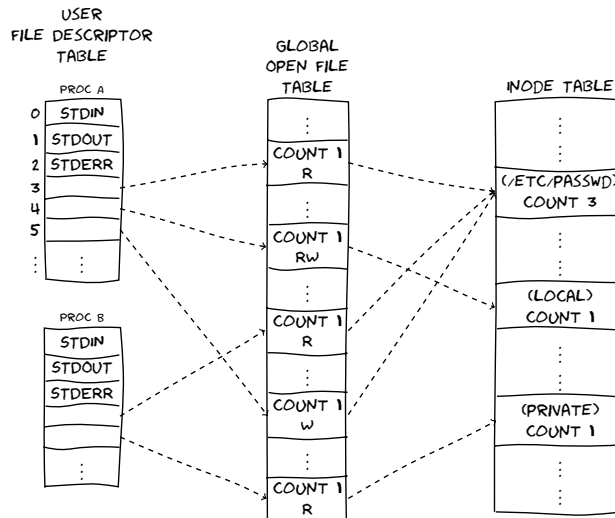
A process executes the following code:

```
fd1 = open("/etc/passwd", O_RDONLY);
fd2 = open("local", O_RDWR);
fd3 = open("/etc/passwd", O_WRONLY);
```



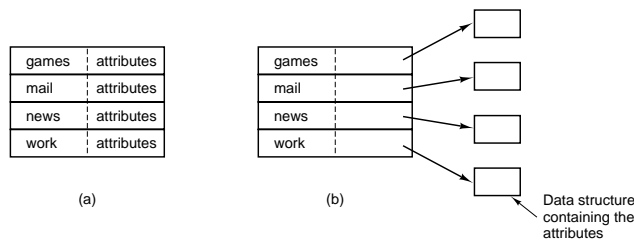
One more process B:

```
fd1 = open("/etc/passwd", O_RDONLY);
fd2 = open("private", O_RDONLY);
```



7 Directories

Implementing Directories



- (a) A simple directory (Windows)
 - fixed size entries
 - disk addresses and attributes in directory entry
- (b) Directory in which each entry just refers to an i-node (UNIX)

Directory entry in glibc

```

1 struct dirent {
2     ino_t      d_ino;        /* Inode number */
3     off_t      d_off;        /* Not an offset; see below */
4     unsigned short d_reclen; /* Length of this record */
5     unsigned char d_type;     /* Type of file; not supported
6                               by all filesystem types */
7     char       d_name[256]; /* Null-terminated filename */
8 };

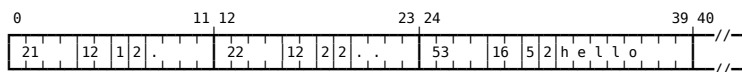
```

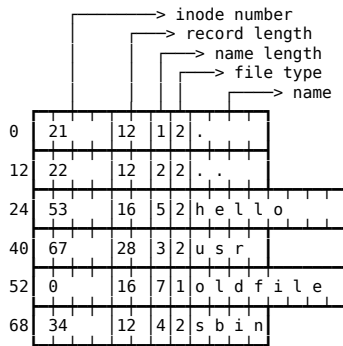
```

$ man readdir
$ view /usr/include/x86_64-linux-gnu/bits/dirent.h

```

Ext2 Directories





- Directories are special files
- “.” and “..” first
- Padding to 4 ×
- inode number is 0 — deleted file

```
ls
1  #include <sys/types.h>
2  #include <dirent.h>
3  #include <stddef.h>
4  #include <stdio.h>
5
6  int main(int argc, char *argv[])
7  {
8      DIR *dp;
9      struct dirent *entry;
10
11     dp = opendir(argv[1]);
12
13     while ( (entry = readdir(dp)) != NULL ){
14         printf("%s\n", entry->d_name);
15     }
16
17     closedir(dp);
18
19     return 0;
20 }
21
22 /* Local Variables: */
23 /* compile-command: "gcc -Wall -Wextra ls.c -o ls" */
24 /* End: */
```

The real ls.c?

- 116 A4 pages
- 5308 lines

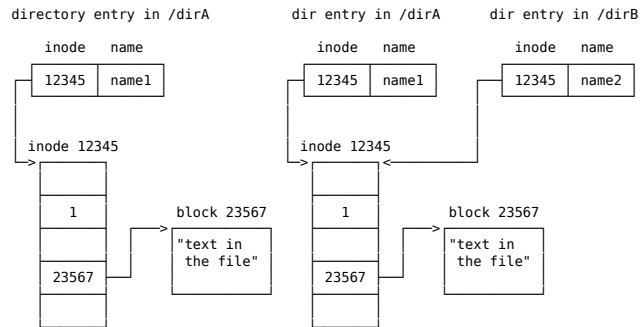
Do one thing, and do it really well.
\$ apt source coreutils

mkdir(), chdir(), rmdir(), getcwd()

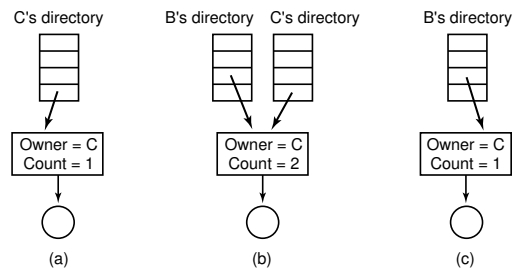
```
1  #include <sys/stat.h>
2  #include <sys/types.h>
3  #include <unistd.h>
4  #include <stdio.h>
5
6  int main(int argc, char *argv[])
7  {
8      char s[100];
9      if ( mkdir(argv[1], S_IRUSR|S_IXUSR) == 0 )
10         chdir(argv[1]);
11     printf("PWD = %s\n", getcwd(s,100));
12     rmdir(argv[1]);
13     return 0;
14 }
15
16 /* Local Variables: */
17 /* compile-command: "gcc -Wall -Wextra mkdir.c -o mkdir" */
18 /* End: */
```

Hard Links

Hard links → the same inode

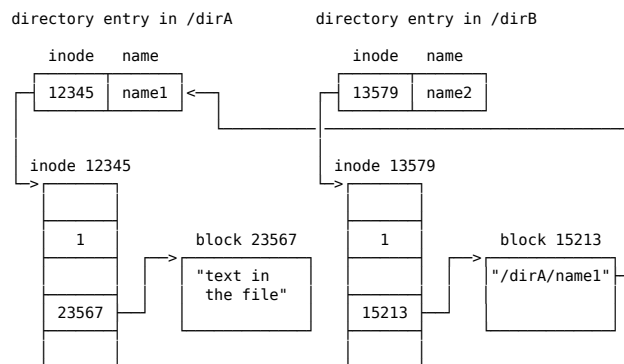


Drawback



Symbolic Links

A symbolic link has its own inode → a directory entry



Fast symbolic link: Short path name (< 60 chars) needs no data block. Can be stored in the 15 pointer fields

`link()`, `unlink()`, `symlink()`

```
1 #include <unistd.h>
2 #include <stdio.h>
3
4 int main(int argc, char *argv[])
5 {
6     /* link(argv[1], argv[2]); */
7     /* symlink(argv[1], argv[2]); */
8     unlink(argv[1]);
9     perror(argv[0]);
10    return 0;
11 }
12
13
```

```
14  /* Local Variables: */  
15  /* compile-command: "gcc -Wall -Wextra link.c -o ln" */  
16  /* End: */
```

Part V

Processes and Threads

8 Virtual Memory

Programs

A program is a file sitting in your hard disk. Two forms:

- Source code, e.g. `hello.c`, human readable
- Executable code, e.g. `a.out`, machine readable

Binary format identification Usually ELF

Machine-language instructions Program algorithm

Entry-point address Where to find `main()`?

Data Initialized variables

Symbol and relocation tables Address of variables, functions...

Shared-library Where to find `printf()`?

More ...

Process

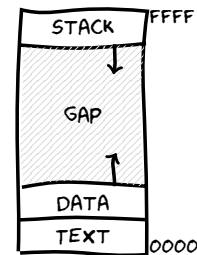
A **process** is an instance of a program in execution

Processes are like human beings:

- ▶ they are generated
- ▶ they have a life
- ▶ they optionally generate one or more child processes, and
- ▶ eventually they die

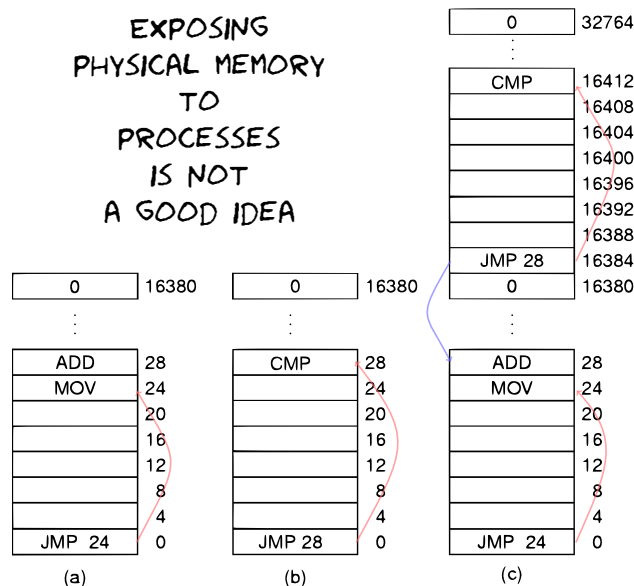
A small difference:

- sex is not really common among processes
- each process has just one parent



Problem With Real Mode

EXPOSING
PHYSICAL MEMORY
TO
PROCESSES
IS NOT
A GOOD IDEA



Protected mode

We need

- Protect the OS from access by user programs
- Protect user programs from one another

Protected mode is an operational mode of x86-compatible CPU.

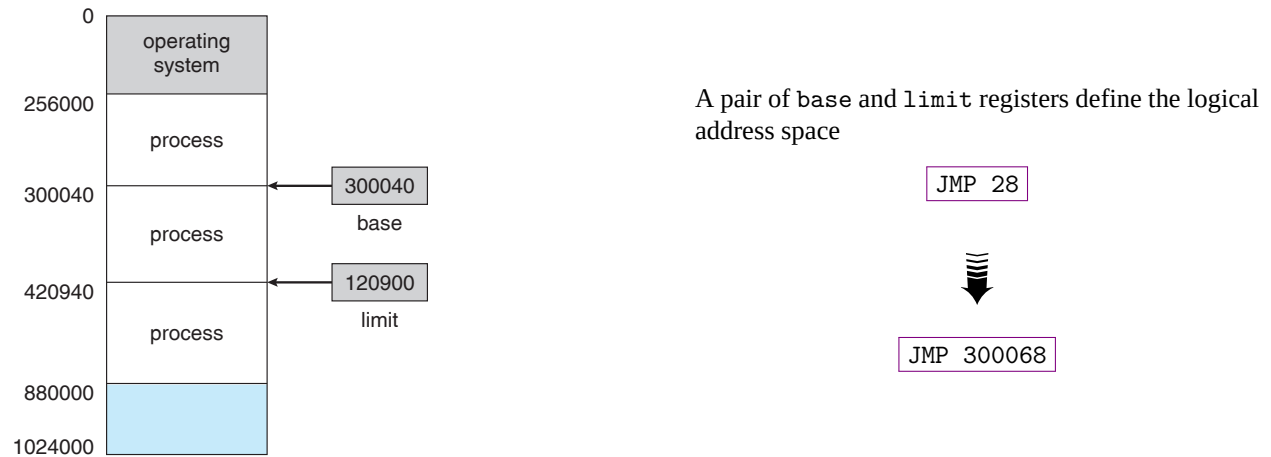
- The purpose is to protect everyone else (including the OS) from your program.

Memory Protection

Logical Address Space

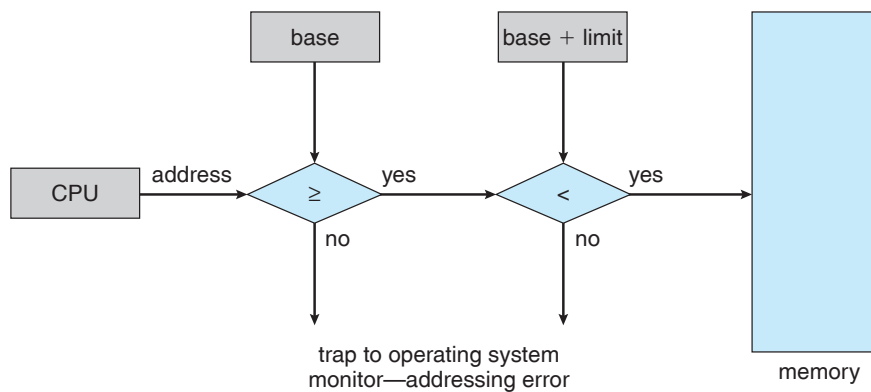
Base register holds the smallest legal physical memory address

Limit register contains the size of the range

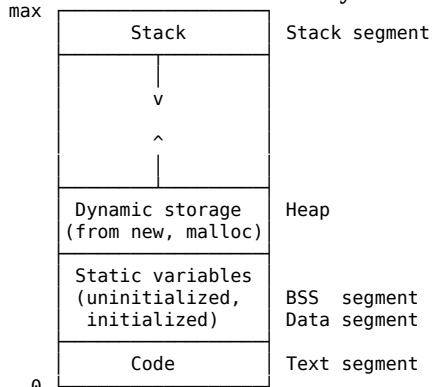


Memory Protection

Base and limit registers



UNIX View of a Process' Memory



THE SIZE OF A PROCESS
(TEXT + DATA + BSS) IS
ESTABLISHED AT COMPILE TIME

text: program code
data: initialized global and static data
bss: uninitialized global and static data
heap: dynamically allocated with malloc, new
stack: local variables

Stack vs. Heap

Stack	Heap
compile-time allocation	run-time allocation
auto clean-up	you clean-up
inflexible	flexible
smaller	bigger
quicker	slower

How large is the ...

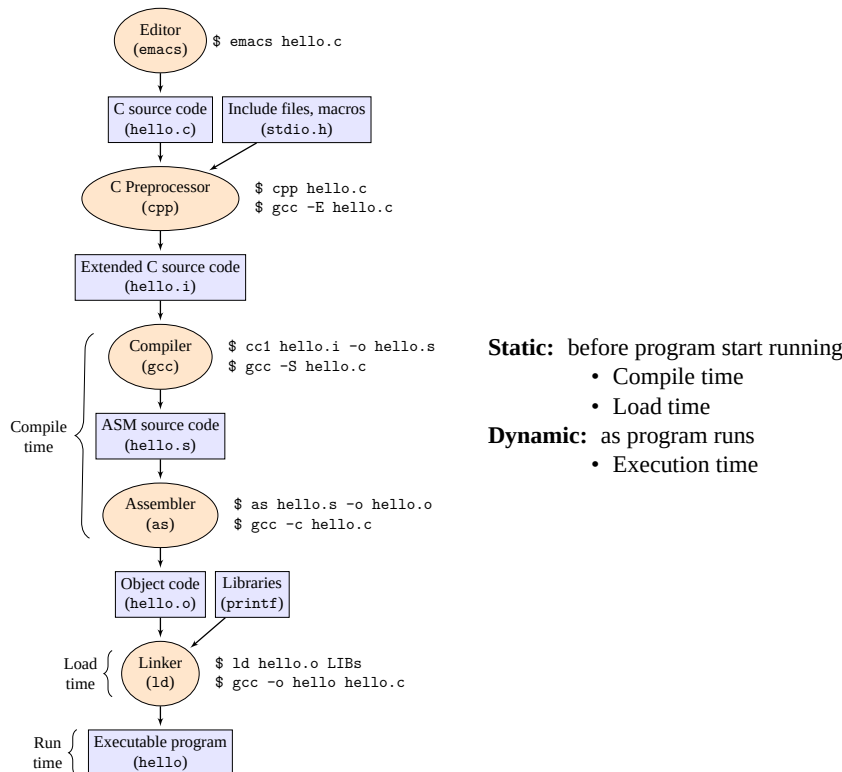
stack: `ulimit -s`

heap: could be as large as your virtual memory

text|data|bss: `size a.out`

Multi-step Processing of a User Program

When is space allocated?



Compiler The name "compiler" is primarily used for programs that translate source code from a high-level programming language to a lower level language (e.g., assembly language or machine code)[*Compiler — Wikipedia, The Free Encyclopedia*].

Assembler An assembler creates object code by translating assembly instruction mnemonics into opcodes, and by resolving symbolic names for memory locations and other entities[*Assembly language — Wikipedia, The Free Encyclopedia*].

Linker Computer programs typically comprise several parts or modules; all these parts/modules need not be contained within a single object file, and in such case refer to each other by means of symbols[*Linker (computing) — Wikipedia, The Free Encyclopedia*].

When a program comprises multiple object files, the linker combines these files into a unified executable program, resolving the symbols as it goes along.

Linkers can take objects from a collection called a library. Some linkers do not include the whole library in the output; they only include its symbols that are referenced from other object files or libraries. Libraries exist for diverse purposes, and one or more system libraries are usually linked in by default.

The linker also takes care of arranging the objects in a program's address space. This may involve relocating code that assumes a specific base address to another base. Since a compiler seldom knows where an object will reside, it often assumes a fixed base location (for example, zero).

Loader An assembler creates object code by translating assembly instruction mnemonics into opcodes, and by resolving symbolic names for memory locations and other entities. ... Loading a program involves reading the contents of executable file, the file containing the program text, into memory, and then carrying out other required preparatory tasks to prepare the executable for running. Once loading is complete, the operating system starts the program by passing control to the loaded program code[*Loader (computing)* — *Wikipedia, The Free Encyclopedia*]

Dynamic linker A dynamic linker is the part of an operating system (OS) that loads (copies from persistent storage to RAM) and links (fills jump tables and relocates pointers) the shared libraries needed by an executable at run time, that is, when it is executed. The specific operating system and executable format determine how the dynamic linker functions and how it is implemented. Linking is often referred to as a process that is performed at compile time of the executable while a dynamic linker is in actuality a special loader that loads external shared libraries into a running process and then binds those shared libraries dynamically to the running process. The specifics of how a dynamic linker functions is operating-system dependent[*Dynamic linker* — *Wikipedia, The Free Encyclopedia*]

Linkers and Loaders allow programs to be built from modules rather than as one big monolith.

See also:

- [Computer Systems: A Programmer's Perspective, Chap. 7, *Linking*].
- COMPILER, ASSEMBLER, LINKER AND LOADER: A BRIEF STORY³.
- Linkers and Loaders⁴.
- [Linkers and Loaders, *Links and loaders*].
- Linux Journal: Linkers and Loaders⁵. Discussing how compilers, links and loaders work and the benefits of shared libraries.

Address Binding

Who assigns memory to segments?

Static-binding: before a program starts running

Compile time: *Compiler* and *assembler* generate an object file for each source file

Load time:

- *Linker* combines all the object files into a single executable object file
- *Loader* (part of OS) loads an executable object file into memory at location(s) determined by the OS
 - invoked via the `execve` system call

Dynamic-binding: as program runs

- Execution time:
 - uses `new` and `malloc` to dynamically allocate memory
 - gets space on stack during function calls
- Address binding has nothing to do with physical memory (RAM). It determines the addresses of objects in the address space (virtual memory) of a process.

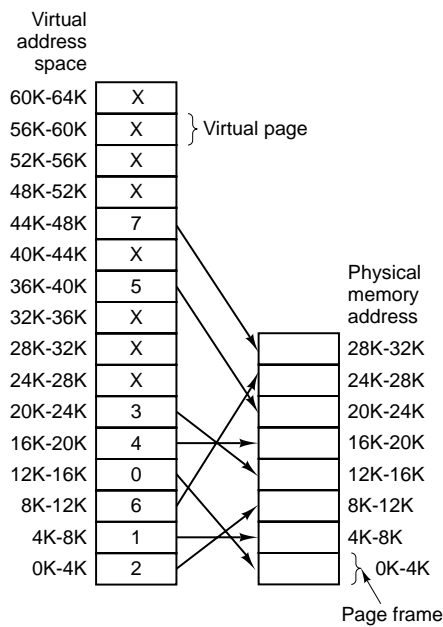
Virtual Memory

Logical memory can be much larger than physical memory

³<http://www.tenouk.com/ModuleW.html>

⁴<http://www.iecc.com/linker/>

⁵<http://www.linuxjournal.com/article/6463>



Address translation

$\text{virtual address} \xrightarrow{\text{page table}} \text{physical address}$

$\text{Page 0} \xrightarrow{\text{map to}} \text{Frame 2}$

$0_{\text{virtual}} \xrightarrow{\text{map to}} 8192_{\text{physical}}$

$20500_{\text{vir}} \xrightarrow{\text{map to}} 12308_{\text{phy}}$
 $(20k + 20)_{\text{vir}} \xrightarrow{\text{map to}} (12k + 20)_{\text{phy}}$

Paging

Address Translation Scheme

Address generated by CPU is divided into:

Page number(p): an index into a page table

Page offset(d): to be copied into memory

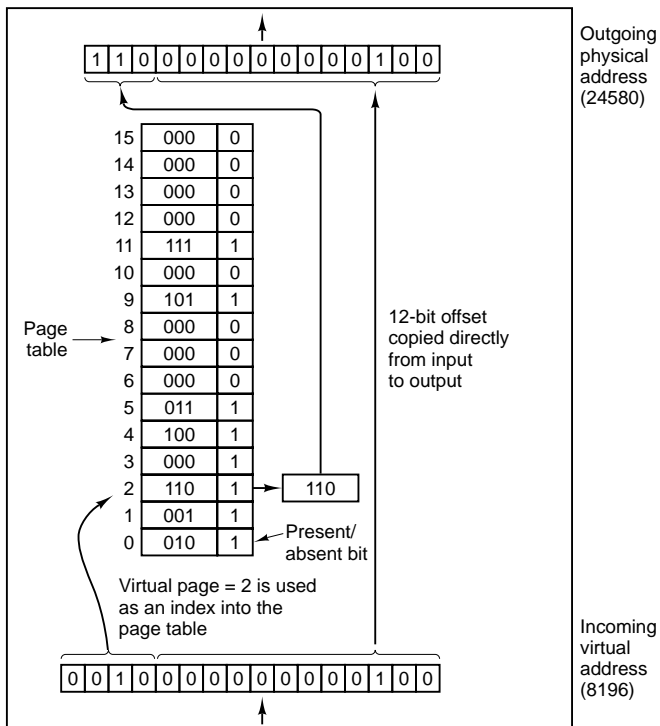
Given *logical address space* (2^m) and *page size* (2^n),

$$\text{number of pages} = \frac{2^m}{2^n} = 2^{m-n}$$

Example: addressing to 001000000000100

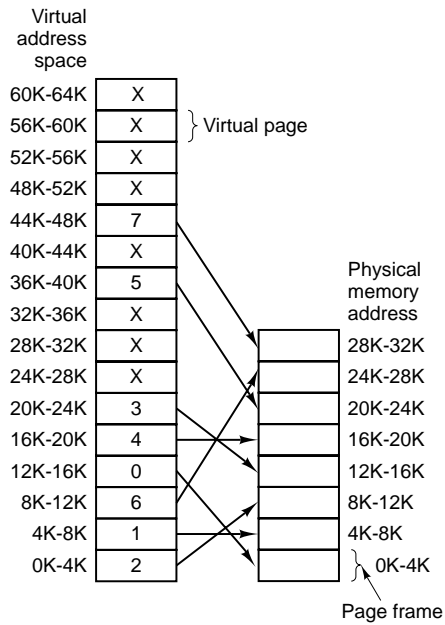
$$\underbrace{\overbrace{0010}^{m-n=4} \overbrace{0000000000100}^{n=12}}_{m=16}$$

page number = 0010 = 2, page offset = 000000000100



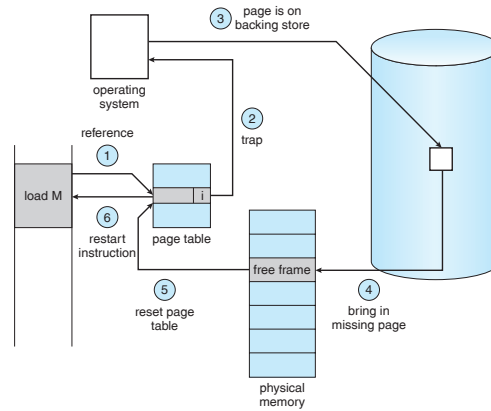
Virtual pages: 16
 Page size: 4k
 Virtual memory: 64K
 Physical frames: 8
 Physical memory: 32K

Page Fault

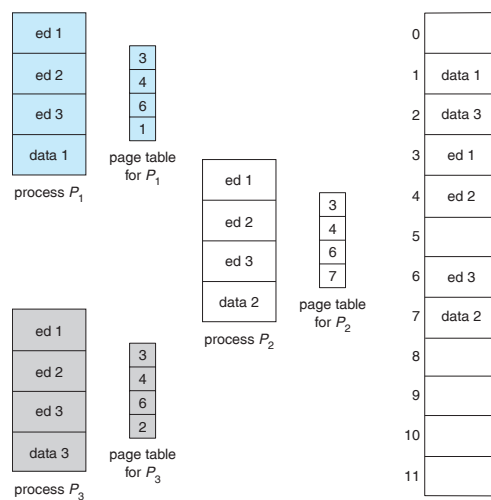


MOV REG, 32780?
 ➡ Page fault & swapping

Page Fault Handling



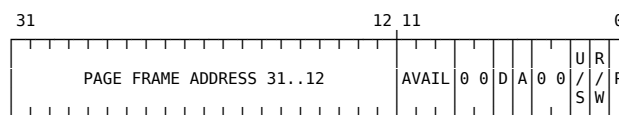
Shared Pages



Page Table Entry

Intel i386 Page Table Entry

- Commonly 4 bytes (32 bits) long
- Page size is usually 4k (2^{12} bytes). OS dependent
\$ getconf PAGESIZE
- Could have $2^{32-12} = 2^{20} = 1M$ pages
Could address $1M \times 4KB = 4GB$ memory



P - PRESENT
R/W - READ/WRITE
U/S - USER/SUPERVISOR
A - ACCESSED
D - DIRTY
AVAIL - AVAILABLE FOR SYSTEMS PROGRAMMER USE

NOTE: 0 INDICATES INTEL RESERVED. DO NOT DEFINE.

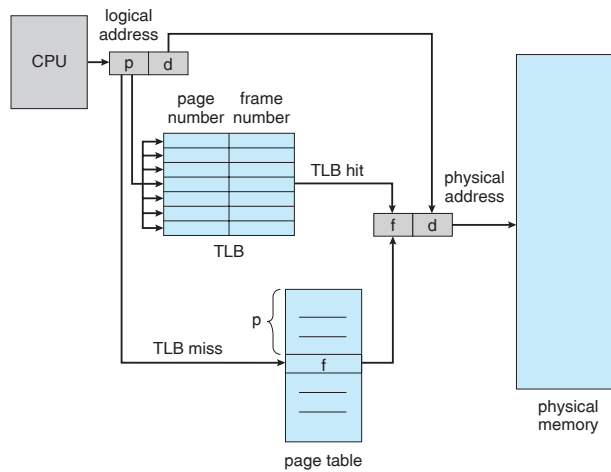
Page Table

- Page table is kept in main memory
- Usually one page table for each process
- *Page-table base register (PTBR)*: A pointer to the page table is stored in PCB
- *Page-table length register (PRLR)*: indicates size of the page table
- Slow

- Requires two memory accesses. One for the page table and one for the data/instruction.
- TLB

Translation Lookaside Buffer (TLB)

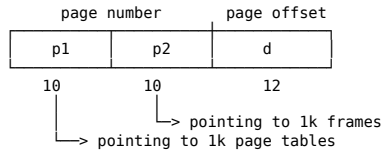
80-20 rule Only a small fraction of the PTEs are heavily read; the rest are barely used at all



Multilevel Page Tables

- a 1M-entry page table eats 4M memory
- while 100 processes running, 400M memory is gone for page tables
- avoid keeping all the page tables in memory all the time

A two-level scheme



p1: is an index into the outer page table

p2: is the displacement within the page of the outer page table

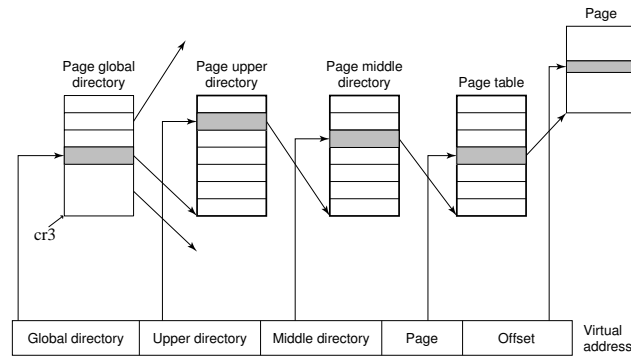
- Split one huge page table into 1k small page tables
 - i.e. the huge page table has 1k entries.
 - Each entry keeps a page frame number of a small page table.
- Each small page table has 1k entries
 - Each entry keeps a page frame number of a physical frame.

Two-Level Page Tables

Example

Paging In Linux

4-level paging for both 32-bit and 64-bit



4-level paging for both 32-bit and 64-bit

- 64-bit: four-level paging
 1. Page Global Directory
 2. Page Upper Directory
 3. Page Middle Directory
 4. Page Table
- 32-bit: two-level paging
 1. Page Global Directory
 2. Page Upper Directory — 0 bits; 1 entry
 3. Page Middle Directory — 0 bits; 1 entry
 4. Page Table

The same code can work on 32-bit and 64-bit architectures

Arch	Page size	Address bits	Paging levels	Address splitting
x86	4KB(12bits)	32	2	10 + 0 + 0 + 10 + 12
x86-PAE	4KB(12bits)	32	3	2 + 0 + 9 + 9 + 12
x86-64	4KB(12bits)	48	4	9 + 9 + 9 + 9 + 12

9 Process

From kernel's point of view

A process consists of


User-space memory program code, variable...

Kernel data structures keep the state of the process

Process Control Block (PCB)

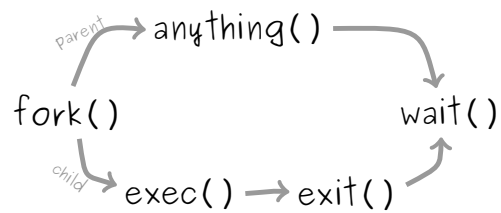
Implementation

A process is *the collection of data structures* that fully describes how far the execution of the program has progressed.

- Each process is represented by a *PCB*
- `task_struct` in 

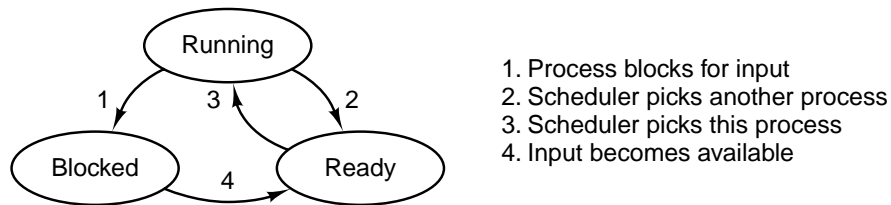
process state
PID
program counter
registers
memory limits
list of open files
...

Process Creation

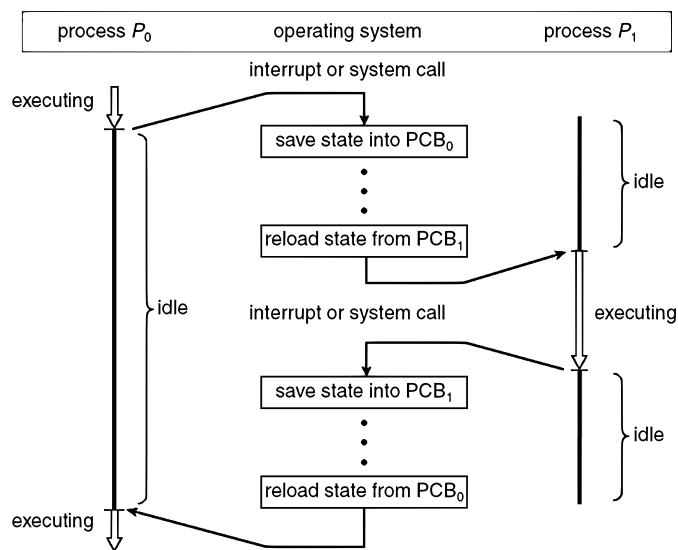


- When a process is created, it is almost identical to its parent
 - It receives a (logical) copy of the parent's address space, and
 - executes the same code as the parent
- The parent and child have separate copies of the data (stack and heap)

Process State Transition



CPU Switch From Process To Process



Forking in C

```

1  #include <stdio.h>
2  #include <unistd.h>
3
4  int main ()
5  {
6      printf("Hello World!\n");
7      fork();
8      printf("Goodbye Cruel World!\n");
9      return 0;
10 }
```

```

1 #include <unistd.h>
2
3 int main(void) {
4     char *argv[] = {NULL};
5     char *envp[] = {NULL};
6     execve("callee.out", argv, envp);
7 }
8
9 /* Local Variables: */
10 /* compile-command: "gcc -Wall
   ↪ -Wextra caller.c -o caller" */
11 /* End: */

```

Fig. 1: caller.c

```

1 #include <stdio.h>
2
3 int main(int argc, char **argv) {
4     if (argc == 0 && argv[0] == NULL)
5         puts("yup");
6 }
7
8 /* Local Variables: */
9 /* compile-command: "gcc -Wall
   ↪ -Wextra callee.c -o callee" */
10 /* End: */

```

Fig. 2: callee.c

```

$ gcc -Wall caller.c -o caller.out
$ gcc -Wall callee.c -o callee.out
$ ./caller.out ./callee.out

```

Fig. 3: argv[0] can be NULL

exec()

```

1 int main()
2 {
3     pid_t pid;
4     /* fork another process */
5     pid = fork();
6     if (pid < 0) { /* error occurred */
7         fprintf(stderr, "Fork Failed");
8         exit(-1);
9     }
10    else if (pid == 0) { /* child process */
11        execlp("/bin/ls", "ls", NULL);
12    }
13    else { /* parent process */
14        /* wait for the child to complete */
15        wait(NULL);
16        printf("Child Complete");
17        exit(0);
18    }
19    return 0;
20 }

```

More about argv[0] int execve(const char *pathname, char *const argv[], char *const envp[]);

- pathname should be the binary image of a program. Or it can be a script (man 2 execve);
- argv[0] is the new process name, usually the same as the basename of pathname, though it can be any other string. It can even be NULL (see Figure 3 for example).

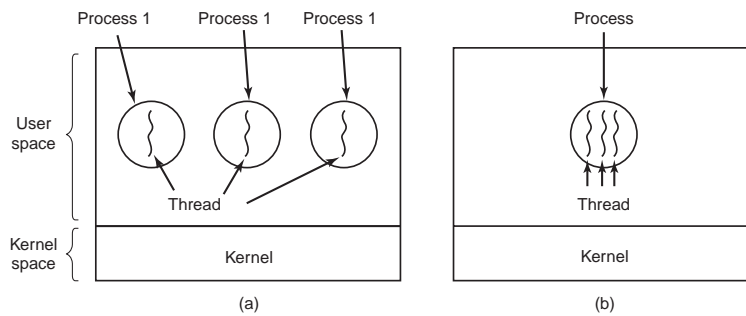
The fact that argv[0] contains the name used to invoke the program can be employed to perform a useful trick. We can create multiple links to (i.e., names for) the same program, and then have the program look at argv[0] and take different actions depending on the name used to invoke it. An example of this technique is provided by the *gzip(1)*, *gunzip(1)*, and *zcat(1)* commands, all of which are links to the same executable file. [The Linux Programming Interface: A Linux and UNIX System Programming Handbook, Sec. 6.6]

- <https://stackoverflow.com/questions/2794150/when-can-argv0-have-null>
- <https://stackoverflow.com/questions/36673765/why-can-the-execve-system-call-run-bin-sh-without-any-argv-arguments-but-not>

10 Thread

Process vs. Thread

a single-threaded process = resource + execution
a multi-threaded process = resource + executions



A process = a unit of resource ownership, used to group resources together;
A thread = a unit of scheduling, scheduled for execution on the CPU.

Threads

code, data, open files, signals...		
thread ID	thread ID	thread ID
program counter	program counter	program counter
register set	register set	register set
stack	stack	stack

POSIX Threads

IEEE 1003.1c The standard for writing portable threaded programs. The threads package it defines is called *Pthreads*, including over 60 function calls, supported by most UNIX systems.

Some of the Pthreads function calls

Thread call	Description
pthread_create	Create a new thread
pthread_exit	Terminate the calling thread
pthread_join	Wait for a specific thread to exit
pthread_yield	Release the CPU to let another thread run
pthread_attr_init	Create and initialize a thread's attribute structure
pthread_attr_destroy	Remove a thread's attribute structure

Pthreads

Example 1

```

1  #include <pthread.h>
2  #include <stdlib.h>
3  #include <unistd.h>
4  #include <stdio.h>
5
6  void *thread_function(void *arg) {
7      int i;
8      for ( i=0; i<10; i++ ) {
9          printf("Thread says hi!, %d\n",i);
10         sleep(1);
11     }

```



```

12     return NULL;
13 }
14
15 int main(void)
16 {
17     pthread_t mythread;
18
19     if( pthread_create(&mythread, NULL, thread_function, NULL) ) {
20         printf("error creating thread.");
21         abort();
22     }
23
24     printf("Can you see my thread working?\n");
25
26     if( pthread_join ( mythread, NULL ) ) {
27         printf("error joining thread.");
28         abort();
29     }
30
31     exit(0);
32 }
33
34 /* Local Variables: */
35 /* compile-command: "gcc -Wall -Wextra thread1.c -o thread1 -pthread" */
36 /* End: */

```

Pthreads

`pthread_t` defined in `pthread.h`, is often called a "thread id" (`tid`);

`pthread_create()` returns zero on success and a non-zero value on failure;

`pthread_join()` returns zero on success and a non-zero value on failure;

How to use pthread?

```

• #include<pthread.h>
$ gcc thread1.c -o thread1 -pthread
$ ./thread1

```

Pthreads

Example 2

```

1  #include <pthread.h>
2  #include <stdio.h>
3  #include <stdlib.h>
4  #include <unistd.h>
5
6  #define NUMBER_OF_THREADS 5
7
8  void *hello(void *tid)
9  {
10     printf ("Hello from thread %d\n", *(int*)tid);
11     pthread_exit(NULL);
12 }
13
14 int main(void)
15 {
16     pthread_t t[NUMBER_OF_THREADS];
17     int status, i;
18
19
20     for (i=0; i<NUMBER_OF_THREADS; i++){
21         printf("Main: creating thread %d ...", i);
22

```

```

23     if( (status = pthread_create(&t[i], NULL, hello, (void *)&i)) ){
24         perror("pthread_create");
25         exit(-1);
26     }
27     puts("done.");
28 }
29
30 for (i=0; i<NUMBER_OF_THREADS; i++){
31     printf("Joining thread %d ...",i);
32
33     if( pthread_join(t[i], NULL) ){
34         perror("pthread_join");
35         abort() ;
36     }
37     puts("done.");
38 }
39 exit(0);
40 }
41
42 /* Local Variables: */
43 /* compile-command: "gcc -Wall -Wextra thread3.c -o thread3 -pthread" */
44 /* End: */

```

Linux Threads

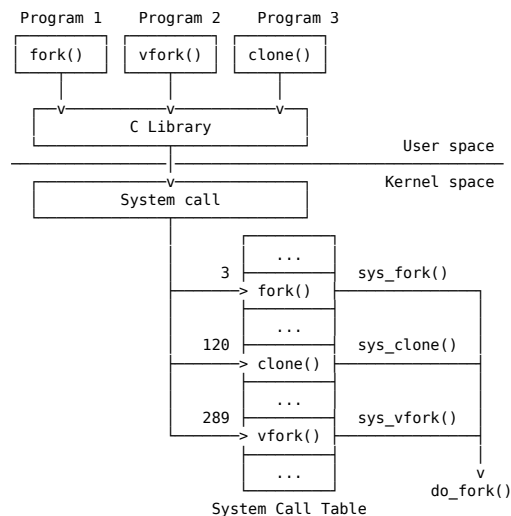
To the Linux kernel, there is no concept of a thread

- Linux implements all threads as standard processes
- To Linux, a thread is merely a process that shares certain resources with other processes
- Some OS (MS Windows, Sun Solaris) have cheap threads and expensive processes.
- Linux processes are already quite lightweight

On a 75MHz Pentium thread: $1.7\mu s$
 fork: $1.8\mu s$

Linux Threads

`clone()` creates a separate process that shares the address space of the calling process. The cloned task behaves *much like* a separate thread.




`clone()`

```

1 #include <sched.h>
2 int clone(int (*fn) (void *), void *child_stack,
3           int flags, void *arg, ...);

```

arg 1 the function to be executed, i.e. `fn(arg)`, which returns an `int`;
arg 2 a pointer  a (usually malloced) memory space to be used as the stack for the new thread;
arg 3 a set of flags used to indicate how much the calling process is to be shared. In fact,
`clone(0) == fork()`
arg 4 the arguments passed to the function.
It returns the PID of the child process or -1 on failure.
\$ man clone

The `clone()` System Call

Some flags:

flag	Shared
CLONE_FS	File-system info
CLONE_VM	Same memory space
CLONE_SIGHAND	Signal handlers
CLONE_FILES	The set of open files

In practice, one should try to avoid calling `clone()` directly

Instead, use a threading library (such as `pthread`) which use `clone()` when starting a thread (such as during a call to `pthread_create()`)

`clone()` Example

```

1  #include <unistd.h>
2  #include <sched.h>
3  #include <sys/types.h>
4  #include <stdlib.h>
5  #include <string.h>
6  #include <stdio.h>
7  #include <fcntl.h>
8
9  int variable;
10
11 int do_something()
12 {
13     variable = 42;
14     _exit(0);
15 }
16
17 int main(void)
18 {
19     void *child_stack;
20     variable = 9;
21
22     child_stack = (void *) malloc(16384);
23     printf("The variable was %d\n", variable);
24
25     clone(do_something, child_stack,
26          CLONE_FS | CLONE_VM | CLONE_FILES, NULL);
27     sleep(1);
28
29     printf("The variable is now %d\n", variable);
30     return 0;
31 }

```

Stack Grows Downwards

```

1 | child_stack = (void**)malloc(8192) + 8192/sizeof(*child_stack);

```

11 Signals

- Signals are software interrupts. Every signal has a name (SIGXXXX). Signals are classic examples of asynchronous events. They occur at what appear to be random times to the process. The process can't simply test a variable (such as `errno`) to see whether a signal has occurred; instead, the process has to tell the kernel "if and when this signal occurs, do the following." [*Advanced programming in the UNIX environment*, chap. 10]
 - Signals are software interrupts sent to a program to indicate that an important event has occurred. The events can vary from user requests to illegal memory access errors. Some signals, such as the interrupt signal, indicate that a user has asked the program to do something that is not in the usual flow of control. (<https://www.tutorialspoint.com/unix/unix-signals-traps>)
 - Signals are similar to interrupts, the difference being that interrupts are mediated by the processor and handled by the kernel while signals are mediated by the kernel (possibly via system calls) and handled by processes. The kernel may pass an interrupt as a signal to the process that caused it (typical examples are SIGSEGV, SIGBUS, SIGILL and SIGFPE). ([https://en.wikipedia.org/wiki/Signal_\(IPC\)](https://en.wikipedia.org/wiki/Signal_(IPC)))
 - `signal(7)`
- ```
$ trap -1
```

### Signals

- Signals are software interrupts
- Every signal has a name (SIGXXXX)
- One process can send a signal to another process

### Sending signals

```
$ [Ctrl]+[c], [Ctrl]+[z], ...
$ kill -signal <pid>
```

### Trapping signals

```
#!/ trap <command> <signals>
```

### Trap

```
1 #!/bin/bash
2
3 sigint(){
4 echo -e "Why Ctrl-c?\n-> "
5 }
6
7 trap sigint SIGINT
8
9 echo -n "-> "
10
11 while read CMD; do
12 $CMD
13 echo -n "-> "
14 done

```

```
#!/ trap "rm -rf $tmpfiles" EXIT
```

### Example

#### SIGINT

```
1 #include <stdio.h>
2 #include <string.h> /* for strlen() */
3 #include <stdlib.h>
4 #include <unistd.h> /* for fork() */
5 #include <sys/wait.h> /* for waitpid() */
6 #include <signal.h>
7
8 #define MAXLINE 4096
```

```

9
10 void sig_int(int signo)
11 {
12 printf("Why Ctrl-c?\n-> ");
13 }
14
15 int main(void)
16 {
17 char buf[MAXLINE];
18 pid_t pid;
19 int status;
20
21 if (signal(SIGINT, sig_int) == SIG_ERR){
22 perror("signal");
23 exit(EXIT_FAILURE);
24 }
25 printf("-> ");
26
27 while(fgets(buf, MAXLINE, stdin) != NULL) {
28 buf[strlen(buf) - 1] = '\0'; /* null */
29
30 if ((pid = fork()) == 0) { /* child */
31 execlp(buf, buf, (char*)0);
32 perror("execlp");
33 exit(127);
34 }
35
36 if((pid = waitpid(pid, &status, 0)) < 0) perror("waitpid");
37 printf("-> ");
38 }
39 exit(EXIT_SUCCESS);
40 }
41
42 /* Local Variables: */
43 /* compile-command: "gcc -Wall -Wextra shell2.c -o shell2" */
44 /* End: */

```

- <https://stackoverflow.com/questions/840501/how-do-function-pointers-in-c-work>

## Example

### SIGUSR1

```

1 #include <signal.h>
2 #include <stdio.h>
3 #include <unistd.h>
4 #include <stdlib.h>
5
6 void sig_usr(int);
7
8 int main(void)
9 {
10 printf("PID = %d\n", getpid());
11
12 if(signal(SIGUSR1, sig_usr) == SIG_ERR){
13 perror("signal<SIGUSR1>");
14 exit(EXIT_FAILURE);
15 }
16
17 for(;;) pause();
18 }
19
20 void sig_usr(int signo)

```

```

21 {
22 if (signo == SIGUSR1)
23 puts("received SIGUSR1.");
24 else{
25 perror("sig_usr");
26 exit(EXIT_FAILURE);
27 }
28 }
29
30 /* Local Variables: */
31 /* compile-command: "gcc -Wall -Wextra sigusr.c -o sigusr" */
32 /* End: */

$ kill -USR1 <PID>

```

## Example

### SIGALRM

```

1 #include <signal.h>
2 #include <stdio.h>
3 #include <unistd.h>
4 #include <stdlib.h>
5
6 void cry(int sig)
7 {
8 puts("C: I'm crying...");
9 kill(getppid(),sig);
10 }
11
12 void complain(int sig)
13 {
14 puts("P: You're noisy.");
15 }
16
17 int main()
18 {
19 if (fork() == 0){
20 signal(SIGALRM, cry);
21 alarm(2);
22 pause();
23 }
24
25 signal(SIGALRM, complain);
26 pause();
27 exit(0);
28 }
29
30 /* Local Variables: */
31 /* compile-command: "gcc -Wall -Wextra alarm.c -o alarm" */
32 /* End: */

```

## Part VI

# Interprocess Communication

- <https://stackoverflow.com/questions/2281204/which-linux-ipc-technique-to-use>
- <https://stackoverflow.com/questions/404604/comparing-unix-linux-ipc>
- <https://www.thegeekstuff.com/2010/08/ipcs-command-examples/>

### Interprocess Communication

#### Example:

```
$ unicode skull | head -1 | cut -f1 -d' ' | sm -
```

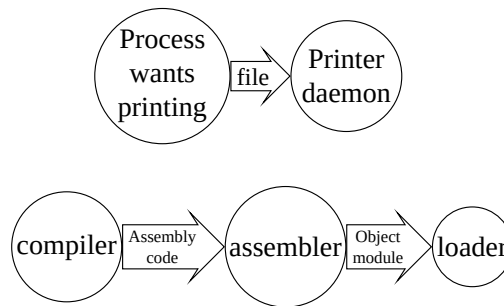
#### IPC issues:

1. How one process can pass information to another
2. Be sure processes do not get into each other's way  
e.g. in an airline reservation system, two processes compete for the last seat
3. Proper sequencing when dependencies are present  
e.g. if A produces data and B prints them, B has to wait until A has produced some data

#### Two models of IPC:

- Shared memory
- Message passing (e.g. sockets)

### Producer-Consumer Problem



### Producer-Consumer Problem

- Consumers don't try to remove objects from Buffer when it is empty.
- Producers don't try to add objects to the Buffer when it is full.

```
1 while(TRUE){
2 while(FULL);
3 item = produceItem();
4 insertItem(item);
5 }

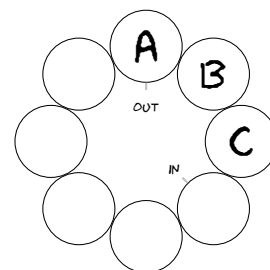
1 while(TRUE){
2 while(EMPTY);
3 item = removeItem();
4 consumeItem(item);
5 }
```

How to define *full/empty*?

### Bounded-Buffer Problem (Circular Array)

**Front(out):** the first full position

**Rear(in):** the next free position



Full or empty when “*front == rear*”?

**Common solution:**

**Full:** when “ $(in + 1) \% BUFFER\_SIZE == out$ ”

Actually, this is “ $full - 1$ ”

**Empty:** when “ $in == out$ ”

Can only use “ $BUFFER\_SIZE - 1$ ” elements

**Shared data:**

```
1 #define BUFFER_SIZE 6
2 typedef struct {
3 ...
4 } item;
5 item buffer[BUFFER_SIZE];
6 int in = 0; //the next free position
7 int out = 0; //the first full position
```

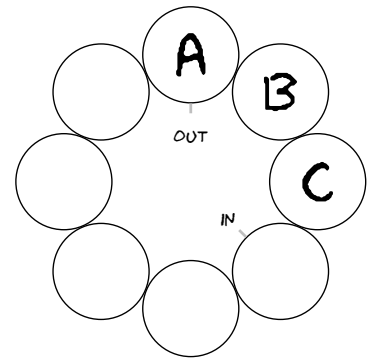
**Bounded-Buffer Problem**

**Producer:**

```
1 while (true) {
2 /* do nothing -- no free buffers */
3 while (((in + 1) % BUFFER_SIZE) == out);
4
5 produce(buffer[in]);
6
7 in = (in + 1) % BUFFER_SIZE;
8 }
```

**Consumer:**

```
1 while (true) {
2 while (in == out); /* do nothing */
3
4 consume(buffer[out]);
5
6 out = (out + 1) % BUFFER_SIZE;
7 }
```

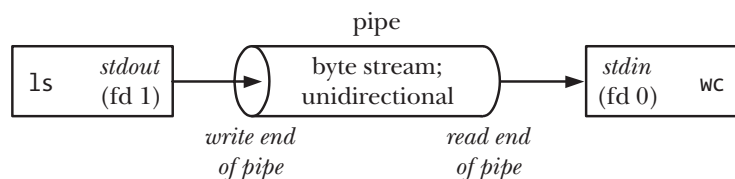


## 12 Pipes and FIFOs

- [The Linux Programming Interface: A Linux and UNIX System Programming Handbook, chap. 44]

**Pipe**

```
$ ls | wc -l
```

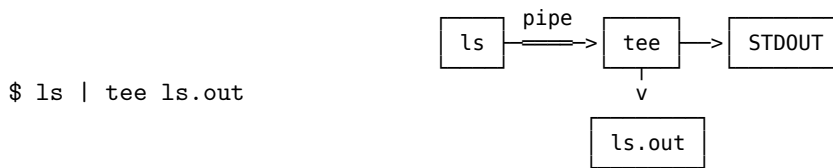


- A pipe is a byte stream
- Unidirectional

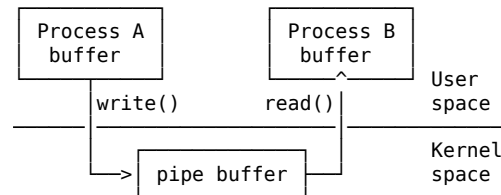


- `read()` would be blocked if nothing written at the other end

tee



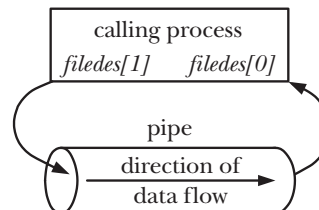
- When we say that a pipe is a byte stream, we mean that there is no concept of messages or message boundaries when using a pipe. The process reading from a pipe can read blocks of data of any size, regardless of the size of blocks written by the writing process. Furthermore, the data passes through the pipe sequentially — bytes are read from a pipe in exactly the order they were written. It is not possible to randomly access the data in a pipe using `lseek()`. [The Linux Programming Interface: A Linux and UNIX System Programming Handbook, chap. 44]



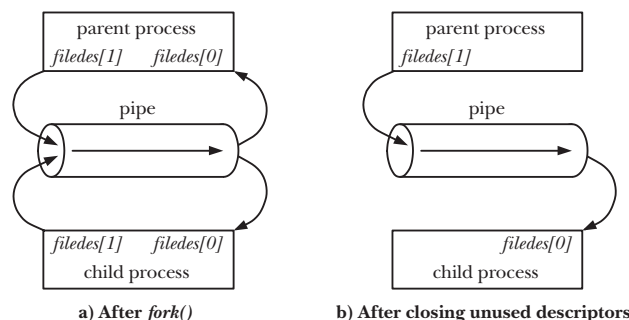
- No direct link between A and B (need system calls)
- A pipe is simply a buffer maintained in kernel memory  
\$ cat /proc/sys/fs/pipe-max-size

`pipe()`

```
1 #include <unistd.h>
2
3 int pipe(int fd[2]);
```



`pipe() + fork()`



- Pipes must have a reader and a writer. If a process tries to write to a pipe that has no reader, it will be sent the SIGPIPE signal from the kernel. This is imperative when more than two processes are involved in a pipeline. (<http://www.tldp.org/LDP/lpg/node20.html>)
- While it is possible for the parent and child to both read from and write to the pipe, this is not usual. Therefore, immediately after the `fork()`, one process closes its descriptor for the write end of the pipe, and the other closes its descriptor for the read end. For example, if the parent is to send data to the child, then it would close its read descriptor for the pipe, `filedes[0]`, while the child would close its write descriptor for the pipe, `filedes[1]`.

One reason that it is not usual to have both the parent and child reading from a single pipe is that if two processes try to simultaneously read from a pipe, we can't be sure which process will be the first to succeed—the two processes race for data. Preventing such races would require the use of some synchronization mechanism. However, if we require bidirectional communication, there is a simpler way: just create two pipes, one for sending data in each direction between the two processes. (If employing this technique, then we need to be wary of deadlocks that may occur if both processes block while trying to read from empty pipes or while trying to write to pipes that are already full.) [*The Linux Programming Interface: A Linux and UNIX System Programming Handbook*, chap. 44, p. 893]

- Pipes can be used for communication between any two (or more) related processes, as long as the pipe was created by a common ancestor before the series of `fork()` calls that led to the existence of the processes. For example, a pipe could be used for communication between a process and its grandchild. The first process creates the pipe, and then forks a child that in turn forks to yield the grandchild. A common scenario is that a pipe is used for communication between two siblings — their parent creates the pipe, and then creates the two children. This is what the shell does when building a pipeline.
- **Closing unused pipe file descriptors.** The process reading from the pipe closes its write descriptor for the pipe, so that, when the other process completes its output and closes its write descriptor, the reader sees end-of-file (once it has read any outstanding data in the pipe). If the reading process doesn't close the write end of the pipe, then, after the other process closes its write descriptor, the reader won't see end-of-file, even after it has read all data from the pipe. Instead, a `read()` would block waiting for data, because the kernel knows that there is still at least one write descriptor open for the pipe. That this descriptor is held open by the reading process itself is irrelevant; in theory, that process could still write to the pipe, even if it is blocked trying to read. For example, the `read()` might be interrupted by a signal handler that writes data to the pipe.

The writing process closes its read descriptor for the pipe for a different reason. When a process tries to write to a pipe for which no process has an open read descriptor, the kernel sends the SIGPIPE signal to the writing process. By default, this signal kills a process. A process can instead arrange to catch or ignore this signal, in which case the `write()` on the pipe fails with the error EPIPE (broken pipe). Receiving the SIGPIPE signal or getting the EPIPE error is a useful indication about the status of the pipe, and this is why unused read descriptors for the pipe should be closed.

If the writing process doesn't close the read end of the pipe, then, even after the other process closes the read end of the pipe, the writing process will still be able to write to the pipe. Eventually, the writing process will fill the pipe, and a further attempt to write will block indefinitely.

One final reason for closing unused file descriptors is that it is only after all file descriptors in all processes that refer to a pipe are closed that the pipe is destroyed and its resources released for reuse by other processes. At this point, any unread data in the pipe is lost.

```

1 #include <sys/wait.h>
2 #include <unistd.h>
3 #include <string.h>
4 #include <stdlib.h>
5 #include <stdio.h>
6
7 #define BUF_SIZE 10
8
9 int main(int argc, char *argv[]) /* Over-simplified! */
10 {
11 int pfd[2]; /* Pipe file descriptors */
12 char buf[BUF_SIZE];
13 ssize_t numRead;
14
15 pipe(pfd); /* Create the pipe */
16
17 switch (fork()) {
18 case 0: /* Child - reads from pipe */
19 close(pfd[1]); /* Write end is unused */
20
21 for(;;) { /* Read data from pipe, echo on stdout */
22 if((numRead = read(pfd[0], buf, BUF_SIZE)) == 0)
23 break; /* End-of-file */
24 if(write(1, buf, numRead) != numRead){

```

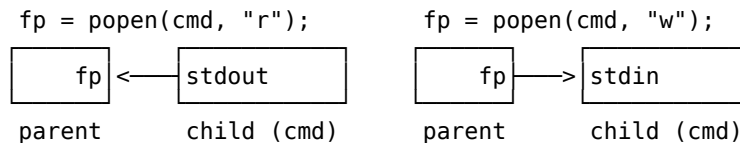
```

25 perror("child - partial/failed write");
26 exit(EXIT_FAILURE);
27 }
28 }
29 puts(""); /* newline */
30
31 close(pfd[0]); _exit(EXIT_SUCCESS);
32
33 default: /* Parent - writes to pipe */
34 close(pfd[0]); /* Read end is unused */
35
36 if((size_t)write(pfd[1], argv[1], strlen(argv[1])) != strlen(argv[1])){
37 perror("parent - partial/failed write");
38 exit(EXIT_FAILURE);
39 }
40
41 close(pfd[1]); /* Child will see EOF */
42
43 wait(NULL); /* Wait for child to finish */
44 exit(EXIT_SUCCESS);
45 }
46 }
47
48 /* Local Variables: */
49 /* compile-command: "gcc -Wall -Wextra simple_pipe.c -o simple-pipe" */
50 /* End: */

```

- <https://stackoverflow.com/questions/5422831/what-is-the-difference-between-using-exit-exit-in-a-conventional-linux-fo>
- `_exit(2)`

## popen()



`popen()` does a `fork()` and `exec()` to execute the `cmd` and returns STD I/O file pointer.

r `fp` is readable (`stdout`)

w `fp` is writable (`stdin`)

- [Advanced programming in the UNIX environment, Sec. 15.3]
- [Beginning linux programming, Sec. 13.2]

## Example

```

1 #include <unistd.h>
2 #include <stdlib.h>
3 #include <stdio.h>
4 #include <string.h>
5
6 int main()
7 {
8 FILE *fp;
9 char buf[1025];
10 int rc;
11
12 memset(buf, '\0', sizeof(buf));
13
14 if((fp = popen("ps ax", "r")) != NULL) {

```

```

15 rc = fread(buf, sizeof(char), 1024, fp);
16 while (rc > 0) {
17 buf[rc - 1] = '\0';
18 printf("Reading %d:-\n %s\n", 1024, buf);
19 rc = fread(buf, sizeof(char), 1024, fp);
20 }
21 pclose(fp);
22 exit(EXIT_SUCCESS);
23 }
24 exit(EXIT_FAILURE);
25 }
26
27
28 /* Local Variables: */
29 /* compile-command: "gcc -Wall -Wextra popen3.c -o popen3" */
30 /* End: */

```

```
$ ps ax | cat
```

## Example

```

1 #include <unistd.h>
2 #include <stdlib.h>
3 #include <stdio.h>
4 #include <string.h>
5
6 int main(int argc, char *argv[])
7 {
8 FILE *fp;
9 char buf[BUFSIZ + 1];
10
11 sprintf(buf, argv[1]);
12
13 if((fp = popen("od -c", "w")) != NULL) {
14 fwrite(buf, sizeof(char), strlen(buf), fp);
15 pclose(fp);
16 exit(EXIT_SUCCESS);
17 }
18 exit(EXIT_FAILURE);
19 }
20
21 /* Local Variables: */
22 /* compile-command: "gcc -Wall -Wextra popen2.c -o popen2" */
23 /* End: */

```

```
$ echo -n hello | od -c
```

## Named Pipe (FIFO)

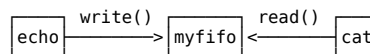
**PIPEs** pass data between related processes.

**FIFOs** pass data between any processes.

```
$ mkfifo myfifo
```

```
$ echo hello > myfifo
```

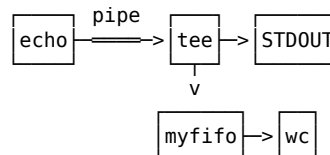
```
$ cat myfifo
```



**tee**

```
$ echo hello | tee myfifo
```

```
$ wc myfifo
```



- [https://en.wikipedia.org/wiki/Named\\_pipe](https://en.wikipedia.org/wiki/Named_pipe)

## IPC With FIFO

```

1 #include <unistd.h>
2 #include <stdlib.h>
3 #include <stdio.h>
4 #include <string.h>
5 #include <fcntl.h>
6 #include <sys/types.h>
7 #include <sys/stat.h>
8
9 #define FIFO_NAME "/tmp/myfifo"
10
11 int main(int argc, char *argv[]) /* Oversimplified */
12 {
13 int fd, i, mode = 0;
14 char c;
15
16 if (argc < 2) {
17 fprintf(stderr, "Usage: %s <O_RDONLY | O_WRONLY | O_NONBLOCK>\n", argv[0]);
18 exit(EXIT_FAILURE);
19 }
20
21 for(i = 1; i < argc; i++) {
22 if (strncmp(++argv, "O_RDONLY", 8) == 0) mode |= O_RDONLY;
23 if (strncmp(*argv, "O_WRONLY", 8) == 0) mode |= O_WRONLY;
24 if (strncmp(*argv, "O_NONBLOCK", 10) == 0) mode |= O_NONBLOCK;
25 }
26
27 if (access(FIFO_NAME, F_OK) == -1) mkfifo(FIFO_NAME, 0777);
28
29 printf("Process %d: FIFO(fd %d, mode %d) opened.\n",
30 getpid(), fd = open(FIFO_NAME, mode), mode);
31
32 if((mode == 0) | (mode == 2048))
33 while(read(fd,&c,1) == 1) putchar(c);
34
35 if((mode == 1) | (mode == 2049))
36 while((c = getchar()) != EOF) write(fd,&c,1);
37
38 exit(EXIT_SUCCESS);
39 }
40
41 /* Local Variables: */
42 /* compile-command: "gcc -Wall -Wextra fifo2.c -o fifo2" */
43 /* End: */

```

```

$ watch 'ls -l /tmp/myfifo'
$./a.out O_RDONLY
$./a.out O_WRONLY
$./a.out O_RDONLY O_NONBLOCK
$./a.out O_WRONLY O_NONBLOCK

```

### O\_NONBLOCK

- A read()/write() will wait on an empty blocking FIFO
- A read() on an empty nonblocking FIFO will return 0 bytes
- open(const char \*path, O\_WRONLY | O\_NONBLOCK);
  - Returns an error (-1) if FIFO not open

- Okay if someone's reading the FIFO
- If opened with O\_RDWR, the result is undefined
- If opened with O\_RDWR, the result is undefined. If you do want to pass data in both directions, it's much better to use a pair of FIFOs or pipes, one for each direction.
- There are four legal combinations of O\_RDONLY, O\_WRONLY, and the O\_NONBLOCK flag.

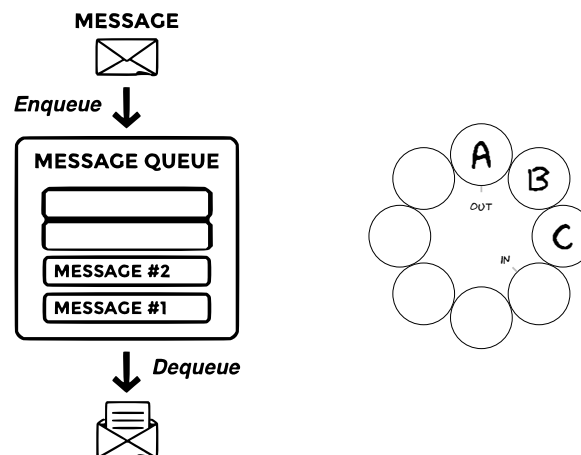
```

1 open(const char *path, O_RDONLY);
2 /* In this case, the open call will block; it will not return until a process opens the
3 same FIFO for writing. */
4
5 open(const char *path, O_RDONLY | O_NONBLOCK);
6 /* The open call will now succeed and return immediately, even if the FIFO has not been
7 opened for writing by any process. */
8
9 open(const char *path, O_WRONLY);
10 /* In this case, the open call will block until a process opens the same FIFO for
11 reading. */
12
13 open(const char *path, O_WRONLY | O_NONBLOCK);
14 /* This will always return immediately, but if no process has the FIFO open for reading,
15 open will return an error, -1, and the FIFO won't be opened. If a process does have the
16 FIFO open for reading, the file descriptor returned can be used for writing to the
17 FIFO. */

```

## 13 Message Queues

### Message Queues



- [The Linux Programming Interface: A Linux and UNIX System Programming Handbook, Sec. 52.3]
- mq\_overview(7)
- sem\_overview(7)
- shm\_overview(7)
- <https://www.uninformativ.de/blog/postings/2016-05-16/0/POSTING-en.html>

### Message Queues

#### Send

```

1 #include <fcntl.h>
2 #include <limits.h>
3 #include <mqqueue.h>
4 #include <stdio.h>
5 #include <stdlib.h>
6 #include <string.h>

```

```

7 #include <sys/stat.h>
8
9 int main(int argc, char **argv)
10 {
11 mqd_t queue;
12 struct mq_attr attrs;
13 size_t msg_len;
14
15 if (argc < 3){
16 fprintf(stderr, "Usage: %s <queuename> <message>\n", argv[0]);
17 return 1;
18 }
19
20 queue = mq_open(argv[1], O_WRONLY | O_CREAT, S_IRUSR | S_IWUSR, NULL);
21 if (queue == (mqd_t)-1){
22 perror("mq_open");
23 return 1;
24 }
25
26 if (mq_getattr(queue, &attrs) == -1){
27 perror("mq_getattr");
28 mq_close(queue);
29 return 1;
30 }
31
32 msg_len = strlen(argv[2]);
33 if (msg_len > LONG_MAX || (long)msg_len > attrs.mq_msgsize){
34 fprintf(stderr, "Your message is too long for the queue.\n");
35 mq_close(queue);
36 return 1;
37 }
38
39 if (mq_send(queue, argv[2], strlen(argv[2]), 0) == -1){
40 perror("mq_send");
41 mq_close(queue);
42 return 1;
43 }
44
45 return 0;
46 }
47
48 /* Local Variables: */
49 /* compile-command: "gcc -Wall -Wextra mq-send.c -o mq-send -lrt" */
50 /* End: */

```

## Message Queues

### Receive

```

1 #include <fcntl.h>
2 #include <mqueue.h>
3 #include <stdio.h>
4 #include <stdlib.h>
5 #include <sys/stat.h>
6
7 int main(int argc, char **argv)
8 {
9 mqd_t queue;
10 struct mq_attr attrs;
11 char *msg_ptr;
12 ssize_t recvd;
13 size_t i;

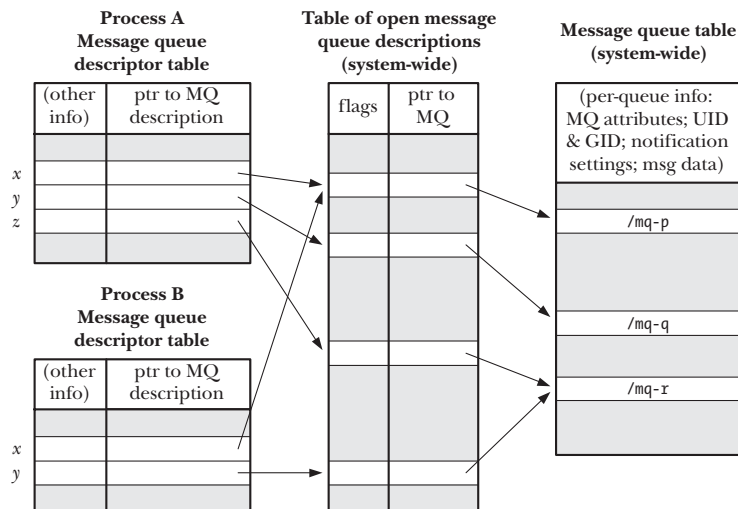
```

```

14
15 if (argc < 2){
16 fprintf(stderr, "Usage: %s <queueenamel>\n", argv[0]);
17 return 1;
18 }
19
20 queue = mq_open(argv[1], O_RDONLY | O_CREAT, S_IRUSR | S_IWUSR, NULL);
21 if (queue == (mqd_t)-1){
22 perror("mq_open");
23 return 1;
24 }
25
26 if (mq_getattr(queue, &attrs) == -1){
27 perror("mq_getattr");
28 mq_close(queue);
29 return 1;
30 }
31
32 msg_ptr = calloc(1, attrs.mq_msgsize);
33 if (msg_ptr == NULL){
34 perror("calloc for msg_ptr");
35 mq_close(queue);
36 return 1;
37 }
38
39 recvd = mq_receive(queue, msg_ptr, attrs.mq_msgsize, NULL);
40 if (recvd == -1){
41 perror("mq_receive");
42 return 1;
43 }
44
45 printf("Message: ");
46 for (i = 0; i < (size_t)recvd; i++)
47 putchar(msg_ptr[i]);
48 puts("");
49 }
50
51 /* Local Variables: */
52 /* compile-command: "gcc -Wall -Wextra mq-recv.c -o mq-recv -lrt" */
53 /* End: */

```

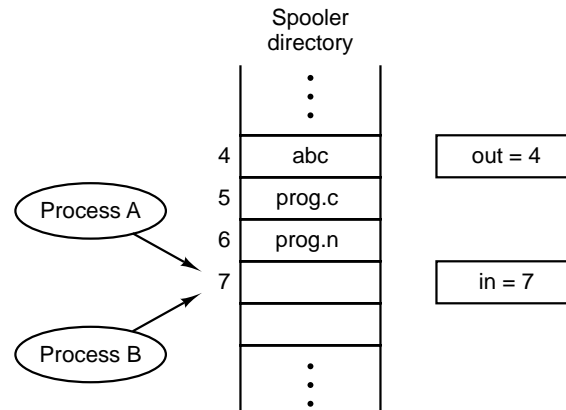
## Relationship Between Kernel Data Structures





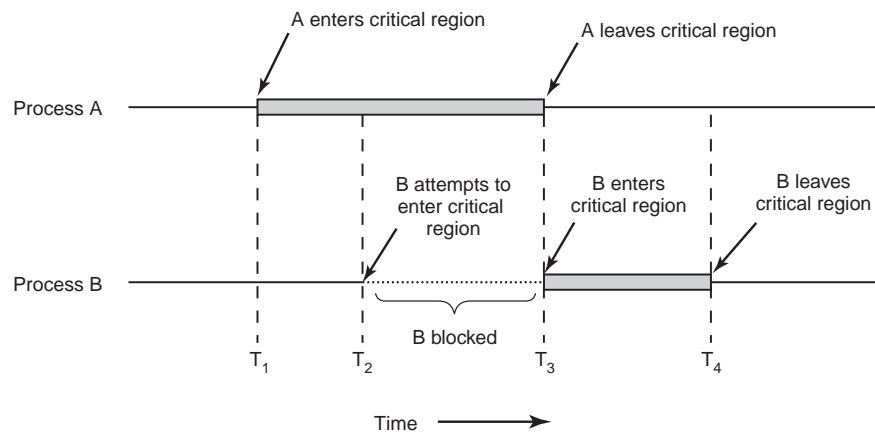
## 14 Semaphores

### Race Conditions



### Mutual Exclusion

**Critical Region** is a piece of code accessing a common resource.



**A solution to the critical region problem must satisfy three conditions**

**Mutual Exclusion:** No two processes may be simultaneously inside their critical regions.

**Progress:** No process running outside its critical region may block other processes.

**Bounded Waiting:** No process should have to wait forever to enter its critical region.

### Mutual Exclusion With Busy Waiting

*Strict Alternation*

```
1 while(TRUE){
2 while(turn != 0);
3 critical_region();
4 turn = 1;
5 noncritical_region();
6 }

1 while(TRUE){
2 while(turn != 1);
3 critical_region();
4 turn = 0;
5 noncritical_region();
6 }
```

⊗ One process can be blocked by another not in its critical region

### Mutual Exclusion With Busy Waiting

*Peterson's Solution*

```

1 | int interest[0] = 0;
2 | int interest[1] = 0;
3 | int turn;

```

**P0**

**P1**

|                                                                                                                                                            |                                                                                                                                                            |
|------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <pre> 1   interest[0] = 1; 2   turn = 1; 3   while(interest[1] == 1 4           &amp;&amp; turn == 1); 5   critical_section(); 6   interest[0] = 0; </pre> | <pre> 1   interest[1] = 1; 2   turn = 0; 3   while(interest[0] == 1 4           &amp;&amp; turn == 0); 5   critical_section(); 6   interest[1] = 0; </pre> |
|------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|

[1] Wikipedia. *Peterson's algorithm* — Wikipedia, *The Free Encyclopedia*. 2015.

## Mutual Exclusion With Busy Waiting

*Lock file*

```

1 | #include <unistd.h>
2 | #include <stdlib.h>
3 | #include <stdio.h>
4 | #include <fcntl.h>
5 | #include <errno.h>
6 |
7 | const char *mylock = "/tmp/LCK.test2";
8 |
9 | int main() {
10 | int fd;
11 |
12 | for(;;){
13 | while((fd = open(mylock, O_RDWR | O_CREAT | O_EXCL, 0444)) != -1) {
14 | printf("Process(%d) - Working in critical region...\n", getpid());
15 | sleep(2); /* working */
16 | close(fd);
17 | if (unlink(mylock) == 0) puts("Done.\nResource unlocked.");
18 | sleep(3); /* non-critical region */
19 | }
20 | printf("Process(%d) - Waiting for lock...\n", getpid());
21 | }
22 |
23 | exit(EXIT_SUCCESS);
24 | }
25 |
26 | /* Local Variables: */
27 | /* compile-command: "gcc -Wall -Wextra lock2.c -o lock2" */
28 | /* End: */

```

⊗ Lock file could be left in system after Ctrl + C

✗

```

1 | #include <unistd.h>
2 | #include <stdlib.h>
3 | #include <stdio.h>
4 | #include <fcntl.h>
5 | #include <errno.h>
6 | #include <signal.h>
7 |
8 | const char *mylock = "/tmp/LCK.test2";
9 |

```

```

10 void sigint(int signo){
11 if (unlink(mylock) == 0) puts("Quit. Lock released.");
12 exit(EXIT_SUCCESS);
13 }
14
15 int main() {
16 int fd;
17
18 signal(SIGINT,sigint);
19
20 for(;;){
21 while((fd = open(mylock, O_RDWR | O_CREAT | O_EXCL, 0444)) != -1) {
22 printf("Process(%d) - Working in critical region...\n", getpid());
23 sleep(2); /* working */
24 close(fd);
25 if (unlink(mylock) == 0) puts("Done.\nResource unlocked.");
26 sleep(3); /* non-critical region */
27 }
28 printf("Process(%d) - Waiting for lock...\n", getpid());
29 }
30 exit(EXIT_SUCCESS);
31 }
32
33 /* Local Variables: */
34 /* compile-command: "gcc -Wall -Wextra lock2-sigint.c -o lock2-sigint" */
35 /* End: */

```

- The `sigint()` function is too crude to be reliable. Need a more sophisticated design.

## What is a Semaphore?

- A locking mechanism
- An integer or ADT

```

1 | down(S){
2 | while(S<=0);
3 | S--;
4 | }
1 | up(S){
2 | S++;
3 | }

```

| Atomic Operations |             |
|-------------------|-------------|
| P()               | V()         |
| Wait()            | Signal()    |
| Down()            | Up()        |
| Decrement()       | Increment() |
| ...               | ...         |

More meaningful names:

- `increment_and_wake_a_waiting_process_if_any()`
- `decrement_and_block_if_the_result_is_negative()`

## Using Semaphore For Signaling

- One thread sends a signal to another to indicate that something has happened
- It solves the serialization problem
  - Signaling makes it possible to guarantee that a section of code in one thread will run before that in another

```

1 | statement a1
2 | sem.signal()
1 | sem.wait()
2 | statement b1

```

What's the initial value of `sem`?

## Example

```
1 #include <stdio.h>
2 #include <unistd.h>
3 #include <stdlib.h>
4 #include <string.h>
5 #include <pthread.h>
6 #include <semaphore.h>
7
8 void *func(void *arg);
9 sem_t sem;
10
11 #define BUFSIZE 1024
12 char buf[BUFSIZE];
13
14 int main() {
15 pthread_t t;
16
17 if(sem_init(&sem, 0, 0) != 0) {
18 perror("Semaphore initialization failed");
19 exit(EXIT_FAILURE);
20 }
21
22 if(pthread_create(&t, NULL, func, NULL) != 0) {
23 perror("Thread creation failed");
24 exit(EXIT_FAILURE);
25 }
26
27 puts("Please input some text. Ctrl-d to quit.");
28
29 while(fgets(buf, BUFSIZE, stdin))
30 sem_post(&sem);
31
32 sem_post(&sem); /* in case of Ctrl-d */
33
34 if(pthread_join(t, NULL) != 0) {
35 perror("Thread join failed");
36 exit(EXIT_FAILURE);
37 }
38
39 sem_destroy(&sem);
40
41 exit(EXIT_SUCCESS);
42 }
43
44 void *func(void *arg) {
45 sem_wait(&sem);
46 while(buf[0] != '\0') {
47 printf("You input %ld characters\n", strlen(buf)-1);
48 buf[0] = '\0'; /* in case of Ctrl-d */
49 sem_wait(&sem);
50 }
51 pthread_exit(NULL);
52 }
53
54 /* Local Variables: */
55 /* compile-command: "gcc -Wall -Wextra thread3.c -o thread3 -pthread" */
56 /* End: */
```

- [Beginning linux programming, Sec. 12.5]
- <https://stackoverflow.com/questions/368322/differences-between-system-v-and-posix-semaphores>

## i++ can go wrong!

```
1 #include <pthread.h>
2 #include <stdlib.h>
3 #include <stdio.h>
4
5 static int glob = 0;
6
7 static void *threadFunc(void *arg) /* loop 'arg' times */
8 {
9 int j;
10
11 for (j = 0; j < *((int *) arg); j++)
12 glob++; /* not atomic! */
13
14 return NULL;
15 }
16
17 int main(int argc, char *argv[])
18 {
19 pthread_t t1, t2;
20 int loops;
21
22 loops = (argc > 1) ? atoi(argv[1]) : 10000000;
23
24 if(pthread_create(&t1, NULL, threadFunc, &loops) != 0){
25 perror("pthread_create 1");
26 }
27
28 if(pthread_create(&t2, NULL, threadFunc, &loops) != 0){
29 perror("pthread_create 2");
30 }
31
32 if(pthread_join(t1, NULL) != 0){
33 perror("pthread_join 1");
34 }
35
36 if(pthread_join(t2, NULL) != 0){
37 perror("pthread_join 2");
38 }
39
40 printf("glob = %d\n", glob);
41 exit(EXIT_SUCCESS);
42 }
43
44 /* Local Variables: */
45 /* compile-command: "gcc -Wall -Wextra atomic-non.c -o atomic-non -pthread" */
46 /* End: */
```

## Atomic

### i++ is not atomic in assembly language

```
1 LOAD [i], r0 ;load the value of 'i' into
2 ;a register from memory
3 ADD r0, 1 ;increment the value
4 ;in the register
5 STORE r0, [i] ;write the updated
 ;value back to memory
```

Interrupts might occur in between. So, i++ needs to be protected with a mutex.

**Mutex** A semaphore that is initialized to 1. In case of:

- 1: A thread may proceed and access the shared variable
- 0: It has to wait for another thread to release the mutex

|                  |                  |
|------------------|------------------|
| 1 mutex.wait()   | 1 mutex.wait()   |
| 2 i++            | 2 i++            |
| 3 mutex.signal() | 3 mutex.signal() |

```
1 #include <pthread.h>
2 #include <stdlib.h>
3 #include <stdio.h>
4
5 static int glob = 0;
6 static pthread_mutex_t mtx = PTHREAD_MUTEX_INITIALIZER;
7
8 static void *threadFunc(void *arg)
9 {
10 int j;
11
12 for (j = 0; j < *((int *) arg); j++) {
13 if (pthread_mutex_lock(&mtx) != 0){
14 perror("pthread_mutex_lock");
15 }
16 glob++;
17 if (pthread_mutex_unlock(&mtx) != 0){
18 perror("pthread_mutex_unlock");
19 }
20 }
21
22 return NULL;
23 }
24
25 int main(int argc, char *argv[])
26 {
27 pthread_t t1, t2;
28 int loops;
29
30 loops = (argc > 1) ? atoi(argv[1]) : 10000000;
31
32 if(pthread_create(&t1, NULL, threadFunc, &loops) != 0){
33 perror("pthread_create");
34 }
35
36 if(pthread_create(&t2, NULL, threadFunc, &loops) != 0){
37 perror("pthread_create");
38 }
39
40 if(pthread_join(t1, NULL) != 0){
41 perror("pthread_join");
42 }
43
44 if(pthread_join(t2, NULL) != 0){
45 perror("pthread_join");
46 }
47
48 printf("glob = %d\n", glob);
49 exit(EXIT_SUCCESS);
50 }
51
52 /* Local Variables: */
```

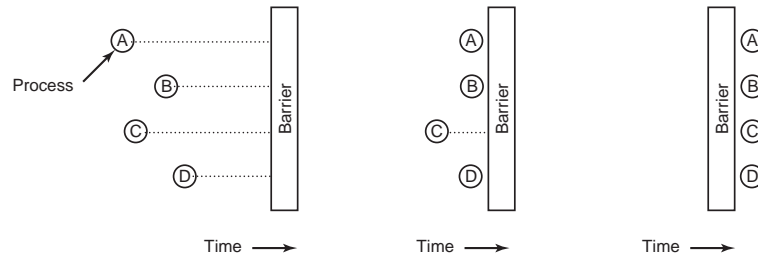
```

53 /* compile-command: "gcc -Wall -Wextra atomic-mutex.c -o atomic-mutex -pthread" */
54 /* End: */

1 #include <semaphore.h>
2 #include <pthread.h>
3 #include <stdio.h>
4 #include <stdlib.h>
5
6 static int glob = 0;
7 static sem_t sem;
8
9 static void *threadFunc(void *arg)
10 {
11 int j;
12
13 for (j = 0; j < *((int *) arg); j++) {
14 if (sem_wait(&sem) == -1){
15 perror("sem_wait");
16 }
17 glob++;
18 if (sem_post(&sem) == -1){
19 perror("sem_post");
20 }
21 }
22
23 return NULL;
24 }
25
26 int main(int argc, char *argv[])
27 {
28 pthread_t t1, t2;
29 int loops;
30
31 loops = (argc > 1) ? atoi(argv[1]) : 10000000;
32
33 if(sem_init(&sem, 0, 1) == -1){
34 perror("sem_init");
35 }
36
37 if(pthread_create(&t1, NULL, threadFunc, &loops) != 0){
38 perror("pthread_create");
39 }
40
41 if(pthread_create(&t2, NULL, threadFunc, &loops) != 0){
42 perror("pthread_create");
43 }
44
45 if(pthread_join(t1, NULL) != 0){
46 perror("pthread_join");
47 }
48
49 if(pthread_join(t2, NULL) != 0){
50 perror("pthread_join");
51 }
52
53 printf("glob = %d\n", glob);
54 exit(EXIT_SUCCESS);
55 }
56
57 /* Local Variables: */
58 /* compile-command: "gcc -Wall -Wextra mutex.c -o mutex -pthread" */
59 /* End: */

```

## Barrier



1. Processes approaching a barrier
2. All processes but one blocked at the barrier
3. When the last process arrives at the barrier, all of them are let through

### Synchronization requirement:

```
specific_task()
critical_point()
```

No thread executes `critical_point()` until after all threads have executed `specific_task()`.

### Barrier Solution

```
1 | n = the number of threads
2 | count = 0
3 | mutex = Semaphore(1)
4 | barrier = Semaphore(0)
```

**count:** keeps track of how many threads have arrived

**mutex:** provides exclusive access to count

**barrier:** is locked ( $\leq 0$ ) until all threads arrive

When `barrier.value < 0`,

`barrier.value == Number of queueing processes`

```
1 | #define CHAIRS 5
2 | semaphore customers = 0;
3 | semaphore bber = ?;
4 | semaphore mutex = 1;
5 | int waiting = 0;
6 |
7 | void barber(void)
8 | {
9 | while (TRUE) {
10 | wait(&customers);
11 | cutHair();
12 | }
13 | }
14 |
15 | void customer(void)
16 | {
17 | if (waiting == CHAIRS)
18 | goHome();
19 | else {
20 | wait(&mutex);
21 | waiting++;
22 | signal(&mutex);
23 | signal(&customers);
24 | wait(&bber);
25 | getHairCut();
26 | signal(&bber);
27 | wait(&mutex);
28 | waiting--;
29 | signal(&mutex);
30 | }
31 | }
```

*Only one thread can pass the barrier!*

### Barrier Solution



```

1 specific_task();
2
3 mutex.wait();
4 count++;
5 mutex.signal();
6
7 if (count == n)
8 barrier.signal();
9
10 barrier.wait();
11 barrier.signal();
12
13 critical_point();

```

```

1 specific_task();
2
3 mutex.wait();
4 count++;
5
6 if (count == n)
7 barrier.signal();
8
9 barrier.wait();
10 barrier.signal();
11 mutex.signal();
12
13 critical_point();

```

💀 Blocking on a semaphore while holding a mutex! 💀

```

barrier.wait();
barrier.signal();

```

## Turnstile

This pattern, a wait and a signal in rapid succession, occurs often enough that it has a name called a *turnstile*, because

- it allows one thread to pass at a time, and
- it can be locked to bar all threads

# 15 Classical IPC Problems

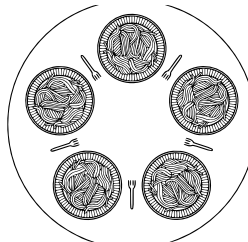
## 15.1 The Dining Philosophers Problem

### The Dining Philosophers Problem

```

1 while True:
2 think()
3 get_forks()
4 eat()
5 put_forks()

```



How to implement `get_forks()` and `put_forks()` to ensure

1. No deadlock
2. No starvation
3. Allow more than one philosopher to eat at the same time

### The Dining Philosophers Problem

#### Deadlock

```

#define N 5 /* number of philosophers */

void philosopher(int i) /* i: philosopher number, from 0 to 4 */
{
 while (TRUE) {
 think(); /* philosopher is thinking */
 take_fork(i); /* take left fork */
 take_fork((i+1) % N); /* take right fork; % is modulo operator */
 eat(); /* yum-yum, spaghetti */
 put_fork(i); /* put left fork back on the table */
 put_fork((i+1) % N); /* put right fork back on the table */
 }
}

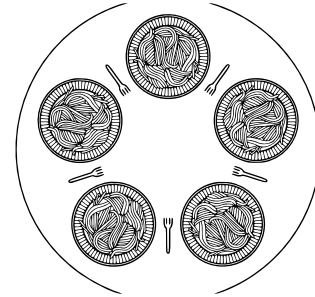
```

- Put down the left fork and wait for a while if the right one is not available? Similar to CSMA/CD — Starvation

## The Dining Philosophers Problem

With One Mutex

```
1 #define N 5
2 semaphore mutex=1;
3
4 void philosopher(int i)
5 {
6 while (TRUE) {
7 think();
8 wait(&mutex);
9 take_fork(i);
10 take_fork((i+1) % N);
11 eat();
12 put_fork(i);
13 put_fork((i+1) % N);
14 signal(&mutex);
15 }
16 }
```



- Only one philosopher can eat at a time.
- How about 2 mutexes? 5 mutexes?

## The Dining Philosophers Problem

AST Solution (Part 1)

A philosopher may only move into eating state if neither neighbor is eating

```
1 #define N 5 /* number of philosophers */
2 #define LEFT (i+N-1)%N /* number of i's left neighbor */
3 #define RIGHT (i+1)%N /* number of i's right neighbor */
4 #define THINKING 0 /* philosopher is thinking */
5 #define HUNGRY 1 /* philosopher is trying to get forks */
6 #define EATING 2 /* philosopher is eating */
7 typedef int semaphore;
8 int state[N]; /* state of everyone */
9 semaphore mutex = 1; /* for critical regions */
10 semaphore s[N]; /* one semaphore per philosopher */
11
12 void philosopher(int i) /* i: philosopher number, from 0 to N-1 */
13 {
14 while (TRUE) {
15 think();
16 take_forks(i); /* acquire two forks or block */
17 eat();
18 put_forks(i); /* put both forks back on table */
19 }
20 }
```

## The Dining Philosophers Problem

AST Solution (Part 2)

```

1 void take_forks(int i) /* i: philosopher number, from 0 to N-1 */
2 {
3 down(&mutex); /* enter critical region */
4 state[i] = HUNGRY; /* record fact that philosopher i is hungry */
5 test(i); /* try to acquire 2 forks */
6 up(&mutex); /* exit critical region */
7 down(&s[i]); /* block if forks were not acquired */
8 }
9 void put_forks(i) /* i: philosopher number, from 0 to N-1 */
10 {
11 down(&mutex); /* enter critical region */
12 state[i] = THINKING; /* philosopher has finished eating */
13 test(LEFT); /* see if left neighbor can now eat */
14 test(RIGHT); /* see if right neighbor can now eat */
15 up(&mutex); /* exit critical region */
16 }
17 void test(i) /* i: philosopher number, from 0 to N-1 */
18 {
19 if (state[i] == HUNGRY && state[LEFT] != EATING && state[RIGHT] != EATING) {
20 state[i] = EATING;
21 up(&s[i]);
22 }
23 }

```

💀 Starvation can happen!

### 15.1.1 Step by step

1. If 5 philosophers take\_forks(i) at the same time, only one can get mutex.
2. The one who gets mutex sets his state to HUNGRY. And then,
3. test(i); try to get 2 forks.
  - If his LEFT and RIGHT are not EATING, success to get 2 forks.
    - i. sets his state to EATING
    - ii. up(&s[i]); The initial value of s(i) is 0.

Now, his LEFT and RIGHT will fail to get 2 forks, even if they could grab mutex.

If either LEFT or RIGHT are EATING, fail to get 2 forks.
4. release mutex
5. down(&s[i]);
  - (a) block if forks are not acquired
  - (b) eat() if 2 forks are acquired
6. After eat()ing, the philosopher doing put\_forks(i) has to get mutex first.
  - because state[i] can be changed by more than one philosopher.
7. After getting mutex, set his state to THINKING
8. test(LEFT); see if LEFT can now eat?
  - If LEFT is HUNGRY, and LEFT's LEFT is not EATING, and LEFT's RIGHT (me) is not EATING
    - i. set LEFT's state to EATING
    - ii. up(&s[LEFT]);

If LEFT is not HUNGRY, or LEFT's LEFT is EATING, or LEFT's RIGHT (me) is EATING, LEFT fails to get 2 forks.
9. test(RIGHT); see if RIGHT can now eat?
10. release mutex

## The Dining Philosophers Problem

### More Solutions

- If there is at least one leftie and at least one rightie, then deadlock is not possible
- Wikipedia: Dining philosophers problem

## 15.2 The Readers-Writers Problem

### The Readers-Writers Problem

**Constraint:** no process may access the shared data for reading or writing while another process is writing to it.

```
1 semaphore mutex = 1;
2 semaphore noOther = 1;
3 int readers = 0;
4
5 void writer(void)
6 {
7 while (TRUE) {
8 wait(&noOther);
9 writing();
10 signal(&noOther);
11 }
12 }
13
14 void reader(void)
15 {
16 while (TRUE) {
17 wait(&mutex);
18 readers++;
19 if (readers == 1)
20 wait(&noOther);
21 signal(&mutex);
22 reading();
23 wait(&mutex);
24 readers--;
25 if (readers == 0)
26 signal(&noOther);
27 signal(&mutex);
28 anything();
29 }
30 }
```

**Starvation** The writer could be blocked forever if there are always someone reading.

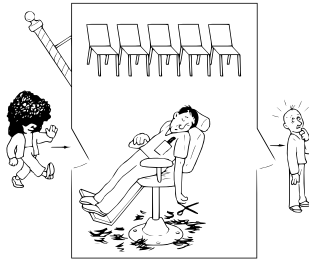
### The Readers-Writers Problem

*No starvation*

```
1 semaphore mutex = 1;
2 semaphore noOther = 1;
3 semaphore turnstile = 1;
4 int readers = 0;
5
6 void writer(void)
7 {
8 while (TRUE) {
9 turnstile.wait();
10 wait(&noOther);
11 writing();
12 signal(&noOther);
13 turnstile.signal();
14 }
15 }
16
17 void reader(void)
18 {
19 while (TRUE) {
20 turnstile.wait();
21 turnstile.signal();
22
23 wait(&mutex);
24 readers++;
25 if (readers == 1)
26 wait(&noOther);
27 signal(&mutex);
28 reading();
29 wait(&mutex);
30 readers--;
31 if (readers == 0)
32 signal(&noOther);
33 signal(&mutex);
34 anything();
35 }
36 }
```

## 15.3 The Sleeping Barber Problem

### The Sleeping Barber Problem



### Where's the problem?

- the barber saw an empty room right before a customer arrives the waiting room;
- Several customer could race for a single chair;

### Solution

```

1 #define CHAIRS 5
2 semaphore customers = 0; // any customers or not?
3 semaphore bber = 0; // barber is busy
4 semaphore mutex = 1;
5 int waiting = 0; // queueing customers

1 void barber(void)
2 {
3 while (TRUE) {
4 wait(&customers);
5 wait(&mutex);
6 waiting--;
7 signal(&mutex);
8 cutHair();
9 signal(&bber);
10 }
11 }

1 void customer(void)
2 {
3 if(waiting == CHAIRS)
4 goHome();
5 else {
6 wait(&mutex);
7 waiting++;
8 signal(&mutex);
9 signal(&customers);
10 wait(&bber);
11 getHairCut();
12 }
13 }

```

### Solution2

```

1 #define CHAIRS 5
2 semaphore customers = 0;
3 semaphore bber = ???;
4 semaphore mutex = 1;
5 int waiting = 0;

1 void customer(void)
2 {
3 wait(&mutex);
4 if (waiting == CHAIRS){
5 signal(&mutex);
6 goHome();
7 } else {
8 waiting++;
9 signal(&customers);
10 signal(&mutex);
11 wait(&bber);
12 getHairCut();
13 wait(&mutex);
14 waiting--;
15 signal(&mutex);
16 signal(&bber);
17 }
18 }

7 void barber(void)
8 {
9 while (TRUE) {
10 wait(&customers);
11 cutHair();
12 }
13 }

```

## 16 Shared Memory

### Write

```

int main(int argc, char *argv[])
{
 int fd;
 size_t len;
 char *addr;
 /* Size of shared memory object */

 if (argc != 3 || strcmp(argv[1], "--help") == 0){
 printf("%s shm-name string\n", argv[0]);
 }

 if ((fd = shm_open(argv[1], O_RDWR | O_CREAT, S_IRUSR | S_IWUSR)) == -1){
 perror("shm_open");
 }

 len = strlen(argv[2]);
 if (ftruncate(fd, len) == -1){ /* Resize object to hold string */
 perror("ftruncate");
 }
 printf("Resized to %ld bytes\n", (long)len);

 addr = mmap(NULL, len, PROT_READ | PROT_WRITE, MAP_SHARED, fd, 0);
 if (addr == MAP_FAILED) perror("mmap");

 if (close(fd) == -1) perror("close");

 printf("copying %ld bytes\n", (long) len);
 memcpy(addr, argv[2], len); /* Copy string to shared memory */
 exit(EXIT_SUCCESS);
}

```

## Read

```

int main(int argc, char *argv[])
{
 int fd;
 char *addr;
 struct stat sb;

 if (argc != 2 || strcmp(argv[1], "--help") == 0){
 printf("%s shm-name\n", argv[0]);
 }

 if ((fd = shm_open(argv[1], O_RDONLY, 0)) == -1){
 perror("shm_open");
 }

 if (fstat(fd, &sb) == -1) perror("fstat"); /* Get object size */

 addr = mmap(NULL, sb.st_size, PROT_READ, MAP_SHARED, fd, 0);
 if (addr == MAP_FAILED) perror("mmap");

 if (close(fd) == -1) perror("close");

 write(STDOUT_FILENO, addr, sb.st_size);
 printf("\n");
 exit(EXIT_SUCCESS);
}

```

## 17 Sockets

### Message Passing

#### Problem with semaphores

- Too low level
- Not suitable for distributed systems

#### Message passing

- No conflicts, easier to implement
- Uses two primitives, send() and receive() system calls:
  - send(destination, &message);
  - receive(source, &message);

## Message Passing

### Design issues

- Message can be lost by network; — ACK
- What if the ACK is lost? — SEQ
- What if two processes have the same name? — socket
- Am I talking with the right guy? Or maybe a MIM? — authentication
- What if the sender and the receiver on the same machine? — Copying messages is always slower than doing a semaphore operation.

## Message Passing

### TCP Header Format

| 0                     |   | 1 |   | 2                |             | 3           |             |             |             |             |             |        |
|-----------------------|---|---|---|------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--------|
| Source Port           |   |   |   | Destination Port |             |             |             |             |             |             |             |        |
| Sequence Number       |   |   |   |                  |             |             |             |             |             |             |             |        |
| Acknowledgment Number |   |   |   |                  |             |             |             |             |             |             |             |        |
| Data Offset           | 0 | 0 | 0 | N<br>S           | C<br>W<br>R | E<br>C<br>R | U<br>A<br>R | P<br>S<br>G | R<br>S<br>K | S<br>S<br>H | F<br>Y<br>T | Window |
| Checksum              |   |   |   | Urgent Pointer   |             |             |             |             |             |             |             |        |
| Options               |   |   |   |                  |             |             |             |             |             |             | Padding     |        |

## Message Passing

### The producer-consumer problem

```
1 #define N 100 /* number of slots in the buffer */
2 void producer(void)
3 {
4 int item;
5 message m; /* message buffer */
6 while (TRUE) {
7 item = produce_item(); /* generate something to put in buffer */
8 receive(consumer, &m); /* wait for an empty to arrive */
9 build_message(&m, item); /* construct a message to send */
10 send(consumer, &m); /* send item to consumer */
11 }
12 }
13
14 void consumer(void)
15 {
16 int item, i;
17 message m;
18 for (i=0; i<N; i++) send(producer, &m); /* send N empties */
19 while (TRUE) {
20 receive(producer, &m); /* get message containing item */
21 item = extract_item(&m); /* extract item from message */
22 send(producer, &m); /* send back empty reply */
23 consume_item(item); /* do something with the item */
24 }
25 }
```

## A TCP Connection

```
wx672@cs3:~$ netstat -at | grep http | grep ESTAB
tcp 0 0 cs3.swfu.edu.cn:http 220.163.96.3:47179 ESTABLISHED
```

| address         | port | address      | port  |
|-----------------|------|--------------|-------|
| cs3.swfu.edu.cn | http | 220.163.96.3 | 47179 |

socket                      socket

a pair of sockets form a TCP connection

## Port numbers

**Port range:** 0 ~ 65535

**Well-known ports:** 0 ~ 1023

|       |       |      |     |        |       |
|-------|-------|------|-----|--------|-------|
| FTP   | 20/21 | SSH  | 22  | Telnet | 23    |
| SMTP  | 25    | DNS  | 53  | DHCP   | 67/68 |
| HTTP  | 80    | POP3 | 110 | HTTPS  | 443   |
| IMAP4 | 143   |      |     |        |       |

## Sockets

To create a socket:

```
fd = socket(domain, type, protocol)
```

**Domain** Determines address format and the range of communication (local or remote). The most commonly used domains are:

| Domain   | Addr structure | Addr format |
|----------|----------------|-------------|
| AF_UNIX  | sockaddr_un    | /path/name  |
| AF_INET  | sockaddr_in    | ip:port     |
| AF_INET6 | sockaddr_in6   | ip6:port    |

**Type** SOCK\_STREAM (🔌), SOCK\_DGRAM (✉)

**Protocol** always 0

## Address Structure

- Different socket domain, different address format, different structure type
- One set of socket syscalls supports all socket domains

```
struct sockaddr {
 sa_family_t sa_family; /* Address family (AF_* constant) */
 char sa_data[14]; /* Socket address (size varies
 according to socket domain) */
};
```

## Example

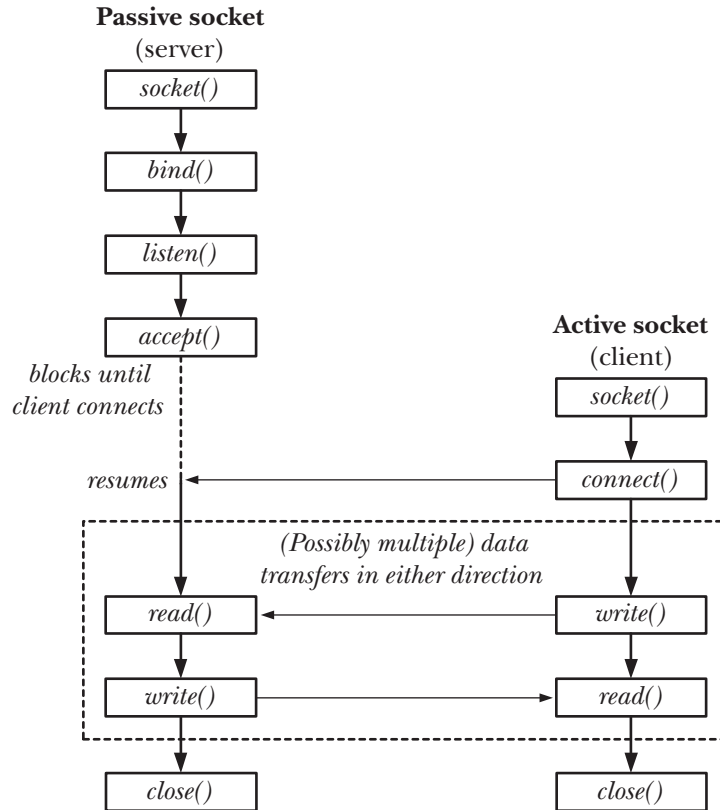
```
struct sockaddr_un addr;
✓ bind(sfd, (struct sockaddr *)&addr, sizeof(struct sockaddr_un));
✗ bind(sfd, &addr, sizeof(struct sockaddr_un));
```

- Why does a cast from `sockaddr_in` to `sockaddr` work? `sockaddr` is defined the way it is to maintain size-compatibility with `sockaddr_in`, which is 16 bytes, and is what `sockaddr` used to be in the days when `AF_INET` was the only address family. But newer families are NOT limited to 14 bytes of data. For instance, `sockaddr_in6` has 26 bytes (the `sin6_addr` field alone is 16 bytes). The ONLY requirement is ALL `sockaddr_*` types MUST start with a 2-byte family member, whose value dictates the size of the remaining data. Presumably, the `bind` function knows what kind of socket descriptor it has as the address family is part of the socket's internal info. `bind()` (and `connect()`) validates the passed in `sockaddr` matches the socket's address family, and converts the `struct sockaddr *` back to a `struct sockaddr_in *`. (For `AF_INET`) after checking that the passed `sockaddr` is the correct family AND byte size (which is why `bind()`, `connect()`, and `accept()` have a parameter to receive the size of the `sockaddr`). If a mismatch occurs, `bind()` (and `connect()` and `accept()`) fails. <https://stackoverflow.com/questions/51287930/why-does-a-cast-from-sockaddr-in-to-sockaddr-work>
- Why do we cast `sockaddr_in` to `sockaddr` when calling `bind()`? <https://stackoverflow.com/questions/21099041/why-do-we-cast-sockaddr-in-to-sockaddr-when-calling-bind>
- `sa_data` is an array of 14 bytes that contains the transport address data. The `SOCKADDR` structure is large enough to contain a transport address for most address families. A WSK application typically does not access the `sa_data` member directly. Instead, a pointer to a `SOCKADDR` structure is normally cast to a pointer to the specific `SOCKADDR` structure type that corresponds to a particular address family. (<https://docs.microsoft.com/en-us/windows/win32/api/ws2def/ns-ws2def-sockaddr>)

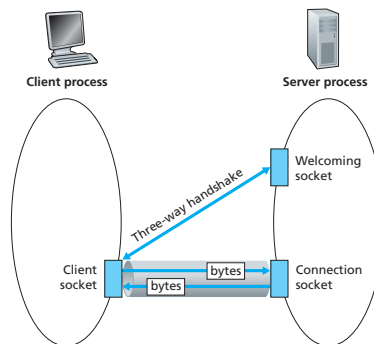


## 17.1 Stream Sockets

### Stream Sockets



### Two Sockets at the Server



### Socket System Calls

`socket()` creates a new socket

`bind()` binds a socket to an address (usually a well-known address on server side)

`listen()` waits for incoming connection requests

`connect()` sends a connection request to peer

`accept()` accepts a connection request

`send()/recv()` data transfer

[*The Linux Programming Interface: A Linux and UNIX System Programming Handbook*, Sec. 56.5]

`int listen(int sockfd, int backlog);` To understand the purpose of the `backlog` argument, we first observe that the client may call `connect()` before the server calls `accept()`. This could happen, for example, because the server is busy handling some other client(s). This results in a pending connection. The kernel must record some information about each pending connection request so that a subsequent `accept()` can be processed.

The backlog argument allows us to limit the number of such pending connections. Connection requests up to this limit succeed immediately. Further connection requests block until a pending connection is accepted (via `accept()`), and thus removed from the queue of pending connections.

`int accept(int sockfd, struct sockaddr *addr, socklen_t *addrlen);` If there are no pending connections when `accept()` is called, the call blocks until a connection request arrives. The key point to understand about `accept()` is that it creates a new socket, and it is this new socket that is connected to the peer socket that performed the `connect()`. A file descriptor for the connected socket is returned as the function result of the `accept()` call. The listening socket (`sockfd`) remains open, and can be used to accept further connections. A typical server application creates one listening socket, binds it to a well-known address, and then handles all client requests by accepting connections via that socket.

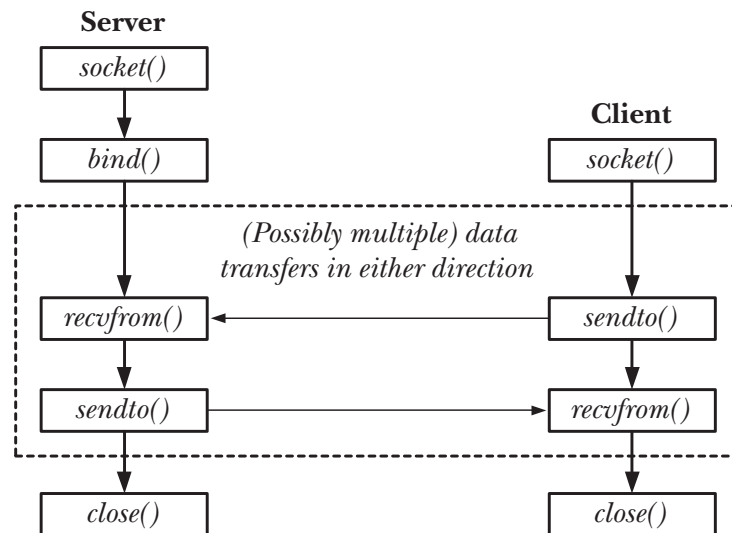
The remaining arguments to `accept()` return the address of the peer socket. The `addr` argument points to a structure that is used to return the socket address. The type of this argument depends on the socket domain (as for `bind()`).

The `addrlen` argument is a value-result argument. It points to an integer that, prior to the call, must be initialized to the size of the buffer pointed to by `addr`, so that the kernel knows how much space is available to return the socket address. Upon return from `accept()`, this integer is set to indicate the number of bytes of data actually copied into the buffer.

If we are not interested in the address of the peer socket, then `addr` and `addrlen` should be specified as `NULL` and 0, respectively.

## 17.2 Datagram Sockets

### Datagram Sockets



## 17.3 Unix Domain Sockets

### Unix Domain Sockets

```

struct sockaddr_un {
 sa_family_t sun_family; /* Always AF_UNIX */
 char sun_path[108]; /* Null-terminated socket pathname */
};

```

### Stream server

```

1 #include <sys/un.h>
2 #include <sys/socket.h>
3 #include <stdio.h>
4 #include <stdlib.h>
5 #include <unistd.h>

```

```

6
7 #define SV_SOCK_PATH "/tmp/us_xfr"
8 #define BUF_SIZE 100
9 #define BACKLOG 5
10
11 int main(void)
12 {
13 struct sockaddr_un addr;
14 int sfd, cfd;
15 ssize_t numRead;
16 char buf[BUF_SIZE];
17
18 if((sfd = socket(AF_UNIX, SOCK_STREAM, 0)) == -1){
19 perror("socket");
20 exit(EXIT_FAILURE);
21 }
22
23 memset(&addr, 0, sizeof(struct sockaddr_un));
24 addr.sun_family = AF_UNIX;
25 strncpy(addr.sun_path, SV_SOCK_PATH, sizeof(addr.sun_path) - 1);
26
27 if(bind(sfd, (struct sockaddr *)&addr, sizeof(struct sockaddr_un)) == -1){
28 perror("bind");
29 exit(EXIT_FAILURE);
30 }
31
32 if(listen(sfd, BACKLOG) == -1){
33 perror("listen");
34 exit(EXIT_FAILURE);
35 }
36
37 for(;;) {
38 if((cfd = accept(sfd, NULL, NULL)) == -1){
39 perror("accept");
40 exit(EXIT_FAILURE);
41 }
42
43 while((numRead = read(cfd, buf, BUF_SIZE)) > 0)
44 if(write(STDOUT_FILENO, buf, numRead) != numRead){
45 perror("partial/failed write");
46 exit(EXIT_FAILURE);
47 }
48
49 if(numRead == -1){
50 perror("read");
51 exit(EXIT_FAILURE);
52 }
53
54 if(close(cfd) == -1){
55 perror("close");
56 exit(EXIT_FAILURE);
57 }
58 }
59 }
60
61 /* Local Variables: */
62 /* compile-command: "gcc -Wall -Wextra us_xfr_sv.c -o us-xfr-sv" */
63 /* End: */

```

## Stream client

```

1 #include <sys/un.h>
2 #include <sys/socket.h>

```

```

3 #include <stdio.h>
4 #include <stdlib.h>
5 #include <unistd.h>
6
7 #define SV_SOCKET_PATH "/tmp/us_xfr"
8 #define BUF_SIZE 100
9
10 int main(void)
11 {
12 struct sockaddr_un addr;
13 int sfd;
14 ssize_t numRead;
15 char buf[BUF_SIZE];
16
17 if((sfd = socket(AF_UNIX, SOCK_STREAM, 0)) == -1){
18 perror("socket");
19 exit(EXIT_FAILURE);
20 }
21
22 memset(&addr, 0, sizeof(struct sockaddr_un));
23 addr.sun_family = AF_UNIX;
24 strncpy(addr.sun_path, SV_SOCKET_PATH, sizeof(addr.sun_path) - 1);
25
26 if (connect(sfd, (struct sockaddr *) &addr, sizeof(struct sockaddr_un)) == -1){
27 perror("connect");
28 exit(EXIT_FAILURE);
29 }
30
31 while ((numRead = read(STDIN_FILENO, buf, BUF_SIZE)) > 0)
32 if (write(sfd, buf, numRead) != numRead){
33 perror("partial/failed write");
34 exit(EXIT_FAILURE);
35 }
36
37 if (numRead == -1){
38 perror("read");
39 exit(EXIT_FAILURE);
40 }
41
42 exit(EXIT_SUCCESS); /* Closes our socket; server sees EOF */
43 }
44
45 /* Local Variables: */
46 /* compile-command: "gcc -Wall -Wextra us_xfr_cl.c -o us-xfr-cl" */
47 /* End: */

```

**Datagram server**

**Datagram client**

## 17.4 Internet Sockets

### Network Byte Order

**Big endian** The most significant byte comes first

**Little endian** The least significant byte comes first

Big endian

| 0x100 0x101 0x102 0x103 |    |    |    |    |     |
|-------------------------|----|----|----|----|-----|
| ...                     | 01 | 23 | 45 | 67 | ... |
| int i = 0x01234567;     |    |    |    |    |     |

Little endian

| 0x100 0x101 0x102 0x103 |    |    |    |    |     |
|-------------------------|----|----|----|----|-----|
| ...                     | 67 | 45 | 23 | 01 | ... |

**Network byte order** is big endian

**Host byte order** Most architectures are big endian. x86 is an exception.

#### Convert int between host and network byte order

```
#include <arpa/inet.h>
uint32_t htonl(uint32_t hostlong);
uint16_t htons(uint16_t hostshort);
uint32_t ntohl(uint32_t netlong);
uint16_t ntohs(uint16_t netshort);
```

Strictly speaking, the use of these four functions is necessary only on systems where the host byte order differs from network byte order. However, these functions should always be used, so that programs are portable to different hardware architectures. On systems where the host byte order is the same as network byte order, these functions simply return their arguments unchanged. [*The Linux Programming Interface: A Linux and UNIX System Programming Handbook*, Sec. 59.2]

## Socket Addresses

### IPv4

```
1 struct in_addr { /* IPv4 4-byte address */
2 in_addr_t s_addr; /* Unsigned 32-bit integer */
3 };
4
5 struct sockaddr_in { /* IPv4 socket address */
6 sa_family_t sin_family; /* Address family (AF_INET) */
7 in_port_t sin_port; /* Port number */
8 struct in_addr sin_addr; /* IPv4 address */
9 unsigned char __pad[X]; /* Pad to size of sockaddr (16 bytes) */
10 };
```

### IPv6

```
1 struct in6_addr { /* IPv6 address structure */
2 uint8_t s6_addr[16]; /* 16 bytes == 128 bits */
3 };
4
5 struct sockaddr_in6 { /* IPv6 socket address */
6 sa_family_t sin6_family; /* Address family (AF_INET6) */
7 in_port_t sin6_port; /* Port number */
8 uint32_t sin6_flowinfo; /* IPv6 flow information */
9 struct in6_addr sin6_addr; /* IPv6 address */
10 uint32_t sin6_scope_id; /* Scope ID (new in kernel 2.4) */
11 };
```