Interprocess Communication

15.1 Introduction

In Chapter 8, we described the process control primitives and saw how to work with multiple processes. But the only way for these processes to exchange information is by passing open files across a fork or an exec or through the file system. We'll now describe other techniques for processes to communicate with one another: interprocess communication (IPC).

In the past, UNIX System IPC was a hodgepodge of various approaches, few of which were portable across all UNIX system implementations. Through the POSIX and The Open Group (formerly X/Open) standardization efforts, the situation has since improved, but differences still exist. Figure 15.1 summarizes the various forms of IPC that are supported by the four implementations discussed in this text.

Note that the Single UNIX Specification (the "SUS" column) allows an implementation to support full-duplex pipes, but requires only half-duplex pipes. An implementation that supports full-duplex pipes will still work with correctly written applications that assume that the underlying operating system supports only half-duplex pipes. We use "(full)" instead of a bullet to show implementations that support half-duplex pipes by using full-duplex pipes.

In Figure 15.1, we show a bullet where basic functionality is supported. For full-duplex pipes, if the feature can be provided through UNIX domain sockets (Section 17.2), we show "UDS" in the column. Some implementations support the feature with pipes and UNIX domain sockets, so these entries have both "UDS" and a bullet

The IPC interfaces introduced as part of the real-time extensions to POSIX.1 were included as options in the Single UNIX Specification. In SUSv4, the semaphore interfaces were moved from an option to the base specification.

IPC type	SUS	FreeBSD 8.0	Linux 3.2.0	Mac OS X 10.6.8	Solaris 10
half-duplex pipes FIFOs	•	(full)	•	•	(full)
full-duplex pipes named full-duplex pipes	allowed obsolescent	•, UDS UDS	UDS UDS	UDS UDS	•, UDS •, UDS
XSI message queues XSI semaphores XSI shared memory	XSI XSI XSI	•	•	•	•
message queues (real-time) semaphores shared memory (real-time)	MSG option • SHM option	•	•	•	•
sockets STREAMS	• obsolescent	•	•	•	•

Figure 15.1 Summary of UNIX System IPC

Named full-duplex pipes are provided as mounted STREAMS-based pipes, but are marked obsolescent in the Single UNIX Specification.

Although support for STREAMS on Linux is available in the "Linux Fast-STREAMS" package from the OpenSS7 project, the package hasn't been updated recently. The latest release of the package from 2008 claims to work with kernels up to Linux 2.6.26.

The first ten forms of IPC in Figure 15.1 are usually restricted to IPC between processes on the same host. The final two rows—sockets and STREAMS—are the only two forms that are generally supported for IPC between processes on different hosts.

We have divided the discussion of IPC into three chapters. In this chapter, we examine classical IPC: pipes, FIFOs, message queues, semaphores, and shared memory. In the next chapter, we take a look at network IPC using the sockets mechanism. In Chapter 17, we take a look at some advanced features of IPC.

15.2 Pipes

Pipes are the oldest form of UNIX System IPC and are provided by all UNIX systems. Pipes have two limitations.

- 1. Historically, they have been half duplex (i.e., data flows in only one direction). Some systems now provide full-duplex pipes, but for maximum portability, we should never assume that this is the case.
- 2. Pipes can be used only between processes that have a common ancestor. Normally, a pipe is created by a process, that process calls fork, and the pipe is used between the parent and the child.

We'll see that FIFOs (Section 15.5) get around the second limitation, and that UNIX domain sockets (Section 17.2) get around both limitations.

Section 15.2 Pipes 535

Despite these limitations, half-duplex pipes are still the most commonly used form of IPC. Every time you type a sequence of commands in a pipeline for the shell to execute, the shell creates a separate process for each command and links the standard output of one process to the standard input of the next using a pipe.

A pipe is created by calling the pipe function.

```
#include <unistd.h> int pipe(int fd[2]); Returns: 0 if OK, -1 on error
```

Two file descriptors are returned through the fd argument: fd[0] is open for reading, and fd[1] is open for writing. The output of fd[1] is the input for fd[0].

Originally in 4.3BSD and 4.4BSD, pipes were implemented using UNIX domain sockets. Even though UNIX domain sockets are full duplex by default, these operating systems hobbled the sockets used with pipes so that they operated in half-duplex mode only.

POSIX.1 allows for implementations to support full-duplex pipes. For these implementations, fd[0] and fd[1] are open for both reading and writing.

Two ways to picture a half-duplex pipe are shown in Figure 15.2. The left half of the figure shows the two ends of the pipe connected in a single process. The right half of the figure emphasizes that the data in the pipe flows through the kernel.

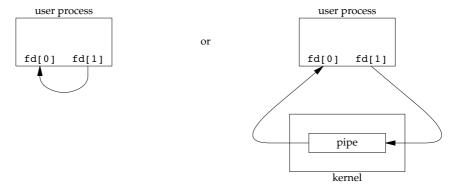


Figure 15.2 Two ways to view a half-duplex pipe

The fstat function (Section 4.2) returns a file type of FIFO for the file descriptor of either end of a pipe. We can test for a pipe with the S_ISFIFO macro.

POSIX.1 states that the st_size member of the stat structure is undefined for pipes. But when the fstat function is applied to the file descriptor for the read end of the pipe, many systems store in st_size the number of bytes available for reading in the pipe. This is, however, nonportable.

A pipe in a single process is next to useless. Normally, the process that calls pipe then calls fork, creating an IPC channel from the parent to the child, or vice versa. Figure 15.3 shows this scenario.

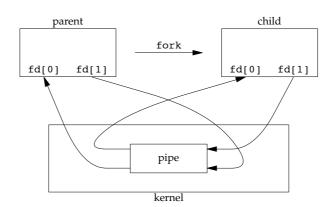


Figure 15.3 Half-duplex pipe after a fork

What happens after the fork depends on which direction of data flow we want. For a pipe from the parent to the child, the parent closes the read end of the pipe (fd[0]), and the child closes the write end (fd[1]). Figure 15.4 shows the resulting arrangement of descriptors.

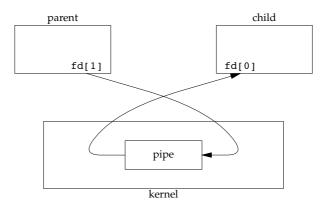


Figure 15.4 Pipe from parent to child

For a pipe from the child to the parent, the parent closes fd[1], and the child closes fd[0].

When one end of a pipe is closed, two rules apply.

1. If we read from a pipe whose write end has been closed, read returns 0 to indicate an end of file after all the data has been read. (Technically, we should say that this end of file is not generated until there are no more writers for the pipe. It's possible to duplicate a pipe descriptor so that multiple processes have the pipe open for writing. Normally, however, there is a single reader and a single writer for a pipe. When we get to FIFOs in the next section, we'll see that often there are multiple writers for a single FIFO.)

Section 15.2 Pipes 537

2. If we write to a pipe whose read end has been closed, the signal SIGPIPE is generated. If we either ignore the signal or catch it and return from the signal handler, write returns -1 with errno set to EPIPE.

When we're writing to a pipe (or FIFO), the constant PIPE_BUF specifies the kernel's pipe buffer size. A write of PIPE_BUF bytes or less will not be interleaved with the writes from other processes to the same pipe (or FIFO). But if multiple processes are writing to a pipe (or FIFO), and if we write more than PIPE_BUF bytes, the data might be interleaved with the data from the other writers. We can determine the value of PIPE_BUF by using pathconf or fpathconf (recall Figure 2.12).

Example

Figure 15.5 shows the code to create a pipe between a parent and its child and to send data down the pipe.

```
#include "apue.h"
main(void)
{
    int
            n;
    int
            fd[2];
    pid t
            pid;
    char
            line[MAXLINE];
    if (pipe(fd) < 0)
        err_sys("pipe error");
    if ((pid = fork()) < 0) {</pre>
        err_sys("fork error");
    } else if (pid > 0) {
                                 /* parent */
        close(fd[0]);
        write(fd[1], "hello world\n", 12);
                                 /* child */
    } else {
        close(fd[1]);
        n = read(fd[0], line, MAXLINE);
        write(STDOUT FILENO, line, n);
    exit(0);
```

Figure 15.5 Send data from parent to child over a pipe

Note that the pipe direction here matches the orientation shown in Figure 15.4.

In the previous example, we called read and write directly on the pipe descriptors. What is more interesting is to duplicate the pipe descriptors onto standard input or standard output. Often, the child then runs some other program, and that program can either read from its standard input (the pipe that we created) or write to its standard output (the pipe).

Example

Consider a program that displays some output that it has created, one page at a time. Rather than reinvent the pagination done by several UNIX system utilities, we want to invoke the user's favorite pager. To avoid writing all the data to a temporary file and calling system to display that file, we want to pipe the output directly to the pager. To do this, we create a pipe, fork a child process, set up the child's standard input to be the read end of the pipe, and exec the user's pager program. Figure 15.6 shows how to do this. (This example takes a command-line argument to specify the name of a file to display. Often, a program of this type would already have the data to display to the terminal in memory.)

```
#include "apue.h"
#include <sys/wait.h>
#define DEF PAGER
                   "/bin/more"
                                    /* default pager program */
int
main(int argc, char *argv[])
    int
           n;
    int
           fd[2];
   pid_t
           pid;
            *pager, *argv0;
   char
   char
           line[MAXLINE];
   FILE
           *fp;
    if (argc != 2)
        err quit("usage: a.out <pathname>");
    if ((fp = fopen(argv[1], "r")) == NULL)
       err sys("can't open %s", argv[1]);
    if (pipe(fd) < 0)
        err_sys("pipe error");
    if ((pid = fork()) < 0) {
        err_sys("fork error");
                                                         /* parent */
    } else if (pid > 0) {
       close(fd[0]);
                            /* close read end */
        /* parent copies argv[1] to pipe */
        while (fgets(line, MAXLINE, fp) != NULL) {
            n = strlen(line);
            if (write(fd[1], line, n) != n)
                err_sys("write error to pipe");
        if (ferror(fp))
            err_sys("fgets error");
        close(fd[1]); /* close write end of pipe for reader */
        if (waitpid(pid, NULL, 0) < 0)
            err_sys("waitpid error");
```

Section 15.2 Pipes 539

```
exit(0);
} else {
                                                 /* child */
   close(fd[1]); /* close write end */
    if (fd[0] != STDIN FILENO) {
        if (dup2(fd[0], STDIN_FILENO) != STDIN FILENO)
            err_sys("dup2 error to stdin");
        close(fd[0]); /* don't need this after dup2 */
    }
    /* get arguments for execl() */
    if ((pager = getenv("PAGER")) == NULL)
       pager = DEF_PAGER;
    if ((argv0 = strrchr(pager, '/')) != NULL)
                       /* step past rightmost slash */
        argv0++;
    else
        argv0 = pager; /* no slash in pager */
    if (execl(pager, argv0, (char *)0) < 0)</pre>
        err sys("execl error for %s", pager);
exit(0);
```

Figure 15.6 Copy file to pager program

Before calling fork, we create a pipe. After the fork, the parent closes its read end, and the child closes its write end. The child then calls dup2 to have its standard input be the read end of the pipe. When the pager program is executed, its standard input will be the read end of the pipe.

When we duplicate one descriptor onto another (fd[0] onto standard input in the child), we have to be careful that the descriptor doesn't already have the desired value. If the descriptor already had the desired value and we called dup2 and close, the single copy of the descriptor would be closed. (Recall the operation of dup2 when its two arguments are equal, discussed in Section 3.12.) In this program, if standard input had not been opened by the shell, the fopen at the beginning of the program should have used descriptor 0, the lowest unused descriptor, so fd[0] should never equal standard input. Nevertheless, whenever we call dup2 and close to duplicate one descriptor onto another, we'll always compare the descriptors first, as a defensive programming measure.

Note how we try to use the environment variable PAGER to obtain the name of the user's pager program. If this doesn't work, we use a default. This is a common usage of environment variables.

Example

Recall the five functions TELL_WAIT, TELL_PARENT, TELL_CHILD, WAIT_PARENT, and WAIT_CHILD from Section 8.9. In Figure 10.24, we showed an implementation using signals. Figure 15.7 shows an implementation using pipes.

```
#include "apue.h"
static int pfd1[2], pfd2[2];
void
TELL_WAIT(void)
    if (pipe(pfd1) < 0 \mid \mid pipe(pfd2) < 0)
        err_sys("pipe error");
}
void
TELL_PARENT(pid_t pid)
{
    if (write(pfd2[1], "c", 1) != 1)
        err_sys("write error");
}
void
WAIT_PARENT(void)
{
    char
            c;
    if (read(pfd1[0], &c, 1) != 1)
        err_sys("read error");
    if (c != 'p')
        err quit("WAIT PARENT: incorrect data");
}
void
TELL_CHILD(pid_t pid)
    if (write(pfd1[1], "p", 1) != 1)
        err_sys("write error");
}
WAIT_CHILD(void)
{
    char
    if (read(pfd2[0], &c, 1) != 1)
        err_sys("read error");
    if (c != 'c')
        err_quit("WAIT_CHILD: incorrect data");
```

Figure 15.7 Routines to let a parent and child synchronize

We create two pipes before the fork, as shown in Figure 15.8. The parent writes the character "p" across the top pipe when TELL_CHILD is called, and the child writes the character "c" across the bottom pipe when TELL_PARENT is called. The corresponding WAIT xxx functions do a blocking read for the single character.

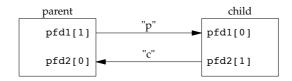


Figure 15.8 Using two pipes for parent-child synchronization

Note that each pipe has an extra reader, which doesn't matter. That is, in addition to the child reading from pfd1[0], the parent has this end of the top pipe open for reading. This doesn't affect us, since the parent doesn't try to read from this pipe.

15.3 popen and pclose Functions

Since a common operation is to create a pipe to another process to either read its output or send it input, the standard I/O library has historically provided the popen and pclose functions. These two functions handle all the dirty work that we've been doing ourselves: creating a pipe, forking a child, closing the unused ends of the pipe, executing a shell to run the command, and waiting for the command to terminate.

The function popen does a fork and exec to execute the *cmdstring* and returns a standard I/O file pointer. If *type* is "r", the file pointer is connected to the standard output of *cmdstring* (Figure 15.9).



Figure 15.9 Result of fp = popen(cmdstring, "r")

If *type* is "w", the file pointer is connected to the standard input of *cmdstring*, as shown in Figure 15.10.



Figure 15.10 Result of fp = popen(cmdstring, "w")

One way to remember the final argument to popen is to remember that, like fopen, the returned file pointer is readable if *type* is "r" or writable if *type* is "w".

The pclose function closes the standard I/O stream, waits for the command to terminate, and returns the termination status of the shell. (We described the termination status in Section 8.6. The system function, described in Section 8.13, also returns the termination status.) If the shell cannot be executed, the termination status returned by pclose is as if the shell had executed exit(127).

The *cmdstring* is executed by the Bourne shell, as in

```
sh -c cmdstring
```

This means that the shell expands any of its special characters in *cmdstring*. This allows us to say, for example,

```
fp = popen("ls *.c", "r");
or
fp = popen("cmd 2>&1", "r");
```

Example

Let's redo the program from Figure 15.6, using popen. This is shown in Figure 15.11.

```
#include "apue.h"
#include <sys/wait.h>
                  "${PAGER:-more}" /* environment variable, or default */
#define PAGER
int
main(int argc, char *argv[])
    char
             line[MAXLINE];
    FILE
             *fpin, *fpout;
    if (argc != 2)
         err_quit("usage: a.out <pathname>");
    if ((fpin = fopen(argv[1], "r")) == NULL)
    err_sys("can't open %s", argv[1]);
    if ((fpout = popen(PAGER, "w")) == NULL)
         err_sys("popen error");
    /* copy argv[1] to pager */
```

```
while (fgets(line, MAXLINE, fpin) != NULL) {
    if (fputs(line, fpout) == EOF)
        err_sys("fputs error to pipe");
}
if (ferror(fpin))
    err_sys("fgets error");
if (pclose(fpout) == -1)
    err_sys("pclose error");
exit(0);
}
```

Figure 15.11 Copy file to pager program using popen

Using popen reduces the amount of code we have to write.

The shell command \${PAGER:-more} says to use the value of the shell variable PAGER if it is defined and non-null; otherwise, use the string more.

Example - popen and pclose Functions

Figure 15.12 shows our version of popen and pclose.

```
#include "apue.h"
#include <errno.h>
#include <fcntl.h>
#include <sys/wait.h>
* Pointer to array allocated at run-time.
*/
static pid_t
                *childpid = NULL;
* From our open_max(), Figure 2.17.
static int
                maxfd;
FILE *
popen(const char *cmdstring, const char *type)
    int
            i;
    int
            pfd[2];
   pid_t
            pid;
   FILE
            *fp;
    /* only allow "r" or "w" */
    if ((type[0] != 'r' && type[0] != 'w') || type[1] != 0) {
        errno = EINVAL;
        return(NULL);
    }
```

```
/* first time through */
if (childpid == NULL) {
    /* allocate zeroed out array for child pids */
   maxfd = open max();
    if ((childpid = calloc(maxfd, sizeof(pid t))) == NULL)
        return(NULL);
}
if (pipe(pfd) < 0)
                  /* errno set by pipe() */
    return(NULL);
if (pfd[0] >= maxfd || pfd[1] >= maxfd) {
   close(pfd[0]);
   close(pfd[1]);
    errno = EMFILE;
    return(NULL);
if ((pid = fork()) < 0) {
    return(NULL); /* errno set by fork() */
} else if (pid == 0) {
                                                /* child */
   if (*type == 'r') {
        close(pfd[0]);
        if (pfd[1] != STDOUT FILENO) {
            dup2(pfd[1], STDOUT_FILENO);
            close(pfd[1]);
        }
    } else {
        close(pfd[1]);
        if (pfd[0] != STDIN_FILENO) {
            dup2(pfd[0], STDIN_FILENO);
            close(pfd[0]);
        }
    }
    /* close all descriptors in childpid[] */
    for (i = 0; i < maxfd; i++)
        if (childpid[i] > 0)
            close(i);
    execl("/bin/sh", "sh", "-c", cmdstring, (char *)0);
    exit(127);
}
/* parent continues... */
if (*type == 'r') {
    close(pfd[1]);
    if ((fp = fdopen(pfd[0], type)) == NULL)
       return(NULL);
} else {
    close(pfd[0]);
    if ((fp = fdopen(pfd[1], type)) == NULL)
        return(NULL);
```

}

```
childpid[fileno(fp)] = pid; /* remember child pid for this fd */
    return(fp);
}
int
pclose(FILE *fp)
            fd, stat;
    pid t
            pid;
    if (childpid == NULL) {
        errno = EINVAL;
                       /* popen() has never been called */
        return(-1);
    }
    fd = fileno(fp);
    if (fd \ge maxfd) {
        errno = EINVAL;
                        /* invalid file descriptor */
        return(-1);
    if ((pid = childpid[fd]) == 0) {
        errno = EINVAL;
                        /* fp wasn't opened by popen() */
        return(-1);
    childpid[fd] = 0;
    if (fclose(fp) == EOF)
        return(-1);
    while (waitpid(pid, &stat, 0) < 0)
        if (errno != EINTR)
            return(-1); /* error other than EINTR from waitpid() */
   return(stat);
                    /* return child's termination status */
}
```

Figure 15.12 The popen and pclose functions

Although the core of popen is similar to the code we've used earlier in this chapter, there are many details that we need to take care of. First, each time popen is called, we have to remember the process ID of the child that we create and either its file descriptor or FILE pointer. We choose to save the child's process ID in the array childpid, which we index by the file descriptor. This way, when pclose is called with the FILE pointer as its argument, we call the standard I/O function fileno to get the file descriptor and then have the child process ID for the call to waitpid. Since it's possible for a given process to call popen more than once, we dynamically allocate the childpid array (the first time popen is called), with room for as many children as there are file descriptors.

Note that our open_max function from Figure 2.17 can return a guess of the maximum number of open files if this value is indeterminate for the system. We need to be careful not to use a pipe file descriptor whose value is larger than (or equal to) what the open_max function returns. In popen, if the value returned by open_max happens to be too small, we close the pipe file descriptors, set errno to EMFILE to indicate too many file descriptors are open, and return -1. In pclose, if the file descriptor corresponding to the file pointer argument is larger than expected, we set errno to EINVAL and return -1.

Calling pipe and fork and then duplicating the appropriate descriptors for each process in the popen function is similar to what we did earlier in this chapter.

POSIX.1 requires that popen close any streams that are still open in the child from previous calls to popen. To do this, we go through the childpid array in the child, closing any descriptors that are still open.

What happens if the caller of pclose has established a signal handler for SIGCHLD? The call to waitpid from pclose would return an error of EINTR. Since the caller is allowed to catch this signal (or any other signal that might interrupt the call to waitpid), we simply call waitpid again if it is interrupted by a caught signal.

Note that if the application calls waitpid and obtains the exit status of the child created by popen, we will call waitpid when the application calls pclose, find that the child no longer exists, and return -1 with errno set to ECHILD. This is the behavior required by POSIX.1 in this situation.

Some early versions of pclose returned an error of EINTR if a signal interrupted the wait. Also, some early versions of pclose blocked or ignored the signals SIGINT, SIGQUIT, and SIGHUP during the wait. This is not allowed by POSIX.1.

Note that popen should never be called by a set-user-ID or set-group-ID program. When it executes the command, popen does the equivalent of

```
execl("/bin/sh", "sh", "-c", command, NULL);
```

which executes the shell and *command* with the environment inherited by the caller. A malicious user can manipulate the environment so that the shell executes commands other than those intended, with the elevated permissions granted by the set-ID file mode.

One thing that popen is especially well suited for is executing simple filters to transform the input or output of the running command. Such is the case when a command wants to build its own pipeline.

Example

Consider an application that writes a prompt to standard output and reads a line from standard input. With the popen function, we can interpose a program between the application and its input to transform the input. Figure 15.13 shows the arrangement of processes in this situation.

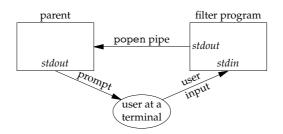


Figure 15.13 Transforming input using popen

The transformation could be pathname expansion, for example, or providing a history mechanism (remembering previously entered commands).

Figure 15.14 shows a simple filter to demonstrate this operation. The filter copies standard input to standard output, converting any uppercase character to lowercase. The reason we're careful to fflush standard output after writing a newline is discussed in the next section when we talk about coprocesses.

Figure 15.14 Filter to convert uppercase characters to lowercase

We compile this filter into the executable file myuclc, which we then invoke from the program in Figure 15.15 using popen.

```
#include "apue.h"
#include <sys/wait.h>
int
main(void)
{
    char line[MAXLINE];
    FILE *fpin;
```

```
if ((fpin = popen("myuclc", "r")) == NULL)
    err_sys("popen error");
for (;;) {
    fputs("prompt> ", stdout);
    fflush(stdout);
    if (fgets(line, MAXLINE, fpin) == NULL) /* read from pipe */
        break;
    if (fputs(line, stdout) == EOF)
        err_sys("fputs error to pipe");
}
if (pclose(fpin) == -1)
    err_sys("pclose error");
putchar('\n');
exit(0);
}
```

Figure 15.15 Invoke uppercase/lowercase filter to read commands

We need to call fflush after writing the prompt, because the standard output is normally line buffered, and the prompt does not contain a newline.

15.4 Coprocesses

A UNIX system filter is a program that reads from standard input and writes to standard output. Filters are normally connected linearly in shell pipelines. A filter becomes a *coprocess* when the same program generates the filter's input and reads the filter's output.

The Korn shell provides coprocesses [Bolsky and Korn 1995]. The Bourne shell, the Bourne-again shell, and the C shell don't provide a way to connect processes together as coprocesses. A coprocess normally runs in the background from a shell, and its standard input and standard output are connected to another program using a pipe. Although the shell syntax required to initiate a coprocess and connect its input and output to other processes is quite contorted (see pp. 62–63 of Bolsky and Korn [1995] for all the details), coprocesses are also useful from a C program.

Whereas popen gives us a one-way pipe to the standard input or from the standard output of another process, with a coprocess we have two one-way pipes to the other process: one to its standard input and one from its standard output. We want to write to its standard input, let it operate on the data, and then read from its standard output.

Example

Let's look at coprocesses with an example. The process creates two pipes: one is the standard input of the coprocess and the other is the standard output of the coprocess. Figure 15.16 shows this arrangement.