7 Alternative I/O Models

7.1 Overview

7.2 Nonblocking I/O

7.3 Signal-driven I/O

7.4 I/O multiplexing: poll()

7.5 Problems with poll() and select()

7.6 The epoll API

7.7 epoll events

7.8 epoll: edge-triggered notification

7.9 epoll: API quirks

7.10 Event-loop programming

Overview

Like select() and poll(), epoll can monitor multiple FDs

epoll returns readiness information in similar manner to poll()

Two main advantages:

epoll provides much better performance when monitoring large numbers of FDs (see TLPI §63.4.5)

epoll provides two notification modes: level-triggered and edge-triggered

Default is level-triggered notification

select() and poll() provide only level-triggered notification

(Signal-driven I/O provides only edge-triggered notification)

Linux-specific, since kernel 2.6.0
**epoll instances**

Central data structure of *epoll* API is an *epoll instance*

- **Persistent** data structure **maintained in kernel space**
  - Referred to in user space via file descriptor
- Can (abstractly) be considered as container for two lists:
  - **Interest list**: list of FDs to be monitored
  - **Ready list**: list of FDs that are ready for I/O
    - Ready list is (dynamic) subset of interest list

**epoll APIs**

The key *epoll* APIs are:

- `epoll_create()`: create a new *epoll* instance and return FD referring to instance
  - FD is used in the calls below
- `epoll_ctl()`: modify interest list of *epoll* instance
  - Add FDs to/remove FDs from interest list
  - Modify events mask for FDs currently in interest list
- `epoll_wait()`: return items from ready list of *epoll* instance
Creating an *epoll* instance: *epoll_create()*

```c
#include <sys/epoll.h>
int epoll_create(int size);
```

- Creates an *epoll* instance
- **size**:
  - Since Linux 2.6.8: serves no purpose, but must be > 0
  - Before Linux 2.6.8: an *estimate* of number of FDs to be monitored via this *epoll* instance
- Returns file descriptor on success, or -1 on error
  - When FD is no longer required, it should be closed via `close()`
- Since Linux 2.6.27, *epoll_create1()* provides improved API
  - See the man page

[TLPI §63.4.1]
Modifying the `epoll` interest list: `epoll_ctl()`

```c
#include <sys/epoll.h>
int epoll_ctl(int epfd, int op, int fd, struct epoll_event *ev);
```

- Modifies the interest list associated with `epoll` FD, `epfd`
- `fd`: identifies which FD in interest list is to have its settings modified
  - Can be FD for pipe, FIFO, terminal, socket, POSIX MQ
  - Can also be an `epoll` FD
    - An `epoll` FD indicates as readable if interest list is nonempty
  - Can't be FD for a regular file or directory

[TLPI §63.4.2]

**epoll_ctl() op argument**

The `epoll_ctl()` *op* argument is one of:

- **EPOLL_CTL_ADD**: add `fd` to interest list
  - `ev` specifies events to be monitored for `fd`
  - If `fd` is already in interest list ⇒ **EEXIST**

- **EPOLL_CTL_MOD**: modify settings of `fd` in interest list
  - `ev` specifies new settings to be associated with `fd`
  - If `fd` is not in interest list ⇒ **ENOENT**

- **EPOLL_CTL_DEL**: remove `fd` from interest list
  - Also removes corresponding entry in ready list, if present
  - `ev` is ignored
  - If `fd` is not in interest list ⇒ **ENOENT**
  - Closing FD automatically removes it from `epoll` interest lists
    - △ But there are exceptions! Manual deletion before `close()` is safer; see later
The *epoll_event* structure

*epoll_ctl()* ev argument is pointer to an *epoll_event* structure:

```c
struct epoll_event {
    uint32_t events;    /* epoll events (bit mask) */
    epoll_data_t data;  /* User data */
};

typedef union epoll_data {
    void     *ptr;    /* Pointer to user-defined data */
    int      fd;      /* File descriptor */
    uint32_t u32;     /* 32-bit integer */
    uint64_t u64;     /* 64-bit integer */
} epoll_data_t;
```

- **ev.events**: bit mask of events to monitor for *fd*
  - (Similar to *events* mask given to *poll()*)
- **data**: info to be passed back to caller of *epoll_wait()* when *fd* later becomes ready
  - **Union field**: value is specified in one of the members

Example: using *epoll_create()* and *epoll_ctl()*

```c
int epfd;
struct epoll_event ev;

epfd = epoll_create(5);

ev.data.fd = fd;
epv.events = EPOLLIN;       /* Monitor for readability */
epoll_ctl(epfd, EPOLL_CTL_ADD, fd, &ev);
```
Waiting for events: `epoll_wait()`

```c
#include <sys/epoll.h>
int epoll_wait(int epfd, struct epoll_event *evlist,
               int maxevents, int timeout);
```

- Returns info about ready FDs in interest list of `epoll` instance of `epfd`
- Blocks until at least one FD is ready
- Info about ready FDs is returned in array `evlist`
  - I.e., can get information about multiple ready FDs with one `epoll_wait()` call
  - (Caller allocates the `evlist` array)
- `maxevents`: size of the `evlist` array
#include <sys/epoll.h>

```c
int epoll_wait(int epfd, struct epoll_event *evlist,
                int maxevents, int timeout);
```

- **timeout** specifies a timeout for call:
  - −1: block until an FD in interest list becomes ready
  - 0: perform a nonblocking “poll” to see if any FDs in interest list are ready
  - > 0: block for up to **timeout** milliseconds or until an FD in interest list becomes ready
    - `epoll_pwait2()` (Linux 5.11) allows nanosecond timeout

- Return value:
  - > 0: number of items placed in **evlist**
  - 0: no FDs became ready within interval specified by **timeout**
  - −1: an error occurred

Info about **multiple** FDs can be returned in the array **evlist**

- Each element of **evlist** returns info about one file descriptor:
  - **events** is a bit mask of events that have occurred for FD
  - **data** is **ev.data** value **currently** associated with FD in the interest list

- **NB**: the FD itself is not returned!
  - Instead, we put FD into **ev.data.fd** when calling `epoll_ctl()`, so that it is returned via `epoll_wait()`
    - (Or, put FD into a structure pointed to by **ev.data.ptr**)

---

Linux/UNIX IPC Programming ©2022, Michael Kerrisk Alternative I/O Models 7-47 §7.7

Linux/UNIX IPC Programming ©2022, Michael Kerrisk Alternative I/O Models 7-48 §7.7
Waiting for events: `epoll_wait()`

```c
#include <sys/epoll.h>
int epoll_wait(int epfd, struct epoll_event *evlist,
               int maxevents, int timeout);
```

- 📢 If `maxevents` FDs are ready, successive `epoll_wait()` calls round-robin through FDs
  - Helps prevent file descriptor starvation
- 📢 In multithreaded programs:
  - One thread can modify interest list (`epoll_ctl()`) while another thread is blocked in `epoll_wait()`
  - `epoll_wait()` call will return if a newly added FD becomes ready

---

### epoll events

Following table shows:

- Bits given in `ev.events` to `epoll_ctl()`
- Bits returned in `evlist[].events` by `epoll_wait()`

<table>
<thead>
<tr>
<th>Bit</th>
<th><code>epoll_ctl()</code>?</th>
<th><code>epoll_wait()</code>?</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPOLLIN</td>
<td>•</td>
<td>•</td>
<td>Normal-priority data can be read</td>
</tr>
<tr>
<td>EPOLLPRI</td>
<td>•</td>
<td>•</td>
<td>High-priority data can be read</td>
</tr>
<tr>
<td>EPOLLRDHUP</td>
<td>•</td>
<td>•</td>
<td>Shutdown on peer socket</td>
</tr>
<tr>
<td>EPOLLOUT</td>
<td>•</td>
<td>•</td>
<td>Data can be written</td>
</tr>
<tr>
<td>EPOLLONESHOT</td>
<td>•</td>
<td>•</td>
<td>Disable monitoring after event notification</td>
</tr>
<tr>
<td>EPOLLET</td>
<td>•</td>
<td>•</td>
<td>Employ edge-triggered notification</td>
</tr>
<tr>
<td>EPOLLERR</td>
<td></td>
<td>•</td>
<td>An error has occurred</td>
</tr>
<tr>
<td>EPOLLHUP</td>
<td></td>
<td>•</td>
<td>A hangup occurred</td>
</tr>
</tbody>
</table>

- Other than `EPOLLONESHOT` and `EPOLLET`, bits have same meaning as similarly named `poll()` bit flags
- `EPOLLIN`, `EPOLLPRI`, `EPOLLRDHUP`, and `EPOLLOUT` are returned by `epoll_wait()` only if specified when adding FD using `epoll_ctl()`
Example: altio/epoll_input.c

Monitors one or more files using `epoll` API to see if input is possible

Suitable files to give as arguments are:
- FIFOs
- Terminal device names
  - (May need to run `sleep` command in FG on the other terminal, to prevent shell stealing input)
- Standard input
  - `/dev/stdin`

Example: altio/epoll_input.c (1)

```c
#define MAX_BUF 1000  /* Max. bytes for read() */
#define MAX_EVENTS 5   /* Max. number of events to be returned from
                     a single epoll_wait() call */

int epfd, ready, fd, s, j, numOpenFds;
struct epoll_event ev;
struct epoll_event evlist[MAX_EVENTS];
char buf[MAX_BUF];

epfd = epoll_create(argc - 1);
```

- Declarations for various variables
- Create an `epoll` instance, obtaining `epoll` FD
Open each of the files named on command line

Each file is monitored for input (EPOLLIN)

$fd$ placed in $\text{ev.data}$, so it is returned by $\text{epoll\_wait()}$

Add the FD to $\text{epoll}$ interest list ($\text{epoll\_ctl()}$)

Track the number of open FDs

Loop, fetching $\text{epoll}$ events and analyzing results

Loop terminates when all FDs has been closed

$\text{epoll\_wait()}$ call places up to $\text{MAX\_EVENTS}$ events in $\text{evlist}$

$\text{timeout} == -1 \Rightarrow$ infinite timeout

Return value of $\text{epoll\_wait()}$ is number of ready FDs
for (j = 0; j < ready; j++) {
    printf(" fd=%d; events: %s%s%s\n", evlist[j].data.fd,
            (evlist[j].events & EPOLLIN) ? "EPOLLIN " : "",
            (evlist[j].events & EPOLLHUP) ? "EPOLLHUP " : "",
            (evlist[j].events & EPOLLERR) ? "EPOLLERR " : "");
    if (evlist[j].events & EPOLLIN) {
        s = read(evlist[j].data.fd, buf, MAX_BUF);
        printf(" read %d bytes: %.*s\n", s, s, buf);
    } else if (evlist[j].events & (EPOLLHUP | EPOLLERR)) {
        printf(" closing fd %d\n", evlist[j].data.fd);
        close(evlist[j].data.fd);
        numOpenFds--;
    }
}

- Scan up to `ready` items in `evlist`
- Display `events` bits
- If `EPOLLIN` event occurred, read some input and display it on `stdout`
  - `%.*s` ⇒ print string with field width taken from argument list (s)
- Otherwise, if error or hangup, close FD and decrements FD count
- Code correctly handles case where both `EPOLLIN` and `EPOLLHUP` are set in `evlist[j].events`

---

**Exercises**

1. Write a client ([template: altio/ex.is_chat_cl.c]) that communicates with the TCP chat server program, `is_chat_sv.c`. The program should be run with the following command line:

   `./is_chat_cl <host> <port> [<nickname>]`

   The program should create a connection to the server, and then use the `epoll` API to monitor both the terminal and the TCP socket for input. All input that becomes available on the socket should be written to the terminal and vice versa.

   - Each time the program sends input from the terminal to the socket, that input should be prepended by the nickname supplied on the command line. If no nickname is supplied, then use the string returned by `getlogin(3)`. (`snprintf(3)` provides an easy way to concatenate the strings.)

   - The program should terminate if it detects end-of-file or an error condition on either file descriptor.

   [Exercise continues on next slide]
Exercises

- Calling `epoll_wait()` with `maxevents==1` will simplify the code!
- As a simplification, you can assume that the socket is always writable (i.e., you don’t need to monitor for the socket for `EPOLLOUT`).
- Bonus points if you find a way to crash the server (reproducibly)!

Exercises

- Write the chat server ([template: altio/ex.is_chat_sv.c]). Note the following points:
  - The program should take one command-line argument: the port number to which it should bind its listening socket.
  - The program should accept and handle multiple simultaneous client connections. Input read from any client should be broadcast to all other clients.
  - Use the `epoll` API to manage the file descriptors.
  - You should use nonblocking file descriptors to ensure that the server never blocks when accepting connections or when reading or writing to clients.
  - When the server detects end-of file or an error (other than `EAGAIN`) while reading or writing on a client connection, it should close that connection. (Remember that closing a file descriptor automatically removes it from any `epoll` interest lists of which it is a member.)
Exercises

Write a program ([template: altio/ex.epoll_pipes.c]) which performs the same task as the altio/poll_pipes.c program, but uses the epoll API instead of poll().

Hints:

- After writing to the pipes, you will need to call epoll_wait() in a loop. The loop should be terminated when epoll_wait() indicates that there are no more ready file descriptors.

- After each call to epoll_wait(), you should display each ready pipe read file descriptor and then drain all input from that file descriptor so that it does not indicate as ready in future calls to epoll_wait().

- In order to drain a pipe without blocking, you will need to make the file descriptor for the read end of the pipe nonblocking.