1 Memory management with threads

In the previous lab, the subjects of creating and joining threads were explored. There are quite a few applications that these ideas could be applied to, but these applications are not exactly the most interesting. Most applications that do interesting things that use threading will most likely have to share variables between threads at some point. This lab will cover the subject of how to share information & variables between threads in a way that produces the desired result.

2 If you like it, then you shoulda put a mutex_lock on it

For the task in the previous lab, there was no reason for variables & information to be shared between multiple threads as each thread computed its own value in the result matrix without need of input from another thread. Though great, it's not a very interesting problem to solve. For more interesting problems, there is a need to share memory between threads. A trivial example of this would be to increment a shared variable between multiple threads. The program to do that is as follows:

```
#include <stdio.h>
#include <stdlib.h>
#include <pthread.h>
#include <errno.h>

void *threadCounter(void *param){
    int *args = (int *)param;
    int i;

    for(i = 0; i < 1000; i++){
        (*args)++;
    }
}

int main(int argc, char** argv){
    pthread_t t1;
    pthread_t t2;
    int shared = 0;

    int err;

    err = pthread_create(&t1, NULL, threadCounter, (void *)&shared);
    if(err != 0){
        errno = err;
        perror("pthread_create");
        exit(1);
    }
    err = pthread_create(&t2, NULL, threadCounter, (void *)&shared);
    if(err != 0){
        errno = err;
        perror("pthread_create");
        exit(1);
    }

    err = pthread_join(t1, NULL);
    if(err != 0){
```

```
        errno = err;
        perror("pthread_join");
        exit(1);
    }
    err = pthread_join(t2, NULL);
    if(err != 0){
        errno = err;
        perror("pthread_join");
        exit(1);
    }

    printf("After both threads are done executing, `shared` = %d\n", shared);
    return 0;
}
```

Save this code snippet as `threaded_count.c`, compile it, and run the program a couple of times, observe the outputs and answer the following questions:

- what is the expected output?
- what is the calculated output?
- what caused the discrepancy between the expected and calculated values?

2.1 Ride into the critical section

One way to avoid the error that occurs in the threaded counter program is for the individual threads to lock the area of code where the accumulation of `shared` is performed, and unlock it once the accumulation is complete for that individual thread. This area is known as the critical section. To maximize performance, it is preferred that the critical section is as small as possible. To perform these locks, the following lines of code are needed:

```
pthread_mutex_t lock;
void *threadFunction(void *args){
.
.
.
pthread_mutex_lock(&lock);
//start of critical section
.
.
.
//end of critical section
pthread_mutex_unlock(&lock);
.
.
.
}
int main(int argc, char** argv){
.
.
.
err = pthread_mutex_init(&lock, NULL);
.
```

```
.
.
err = pthread_mutex_destroy(&lock);
return 0;
}
```

For more information regarding the init, lock, and unlock calls, consult `man 3 pthread_mutex_init`, `man 3 pthread_mutex_lock`, and `man 3 pthread_mutex_unlock` respectively. As seen above, the variable `lock` is declared as a global variable so that all the threads can get access to it. It is initialized in the main thread using `pthread_mutex_init`, and the threads use it to lock critical sections using `pthread_mutex_lock`. Once the critical section is complete, the critical section is unlocked using `pthread_mutex_unlock`. Finally, before the program exits, destroy the mutex using the `pthread_mutex_destroy` function call.

Now, add the mutex and the calls to `pthread_mutex_lock` and `pthread_mutex_unlock` to the counting thread functions in `threaded_count.c`, compile it, run it, and answer the following questions:

- Did this fix the issue with the original code?

3 See the threads in the streets, with not enough to do

Using mutexes to lock and unlock are great for avoiding race conditions, like what was shown in `threaded_count.c`, but it doesn't do a very good job of keeping threads from executing when we do not want them to.

Suppose that a program has one producer thread P, and two consumer threads C1 and C2. To ensure correctness of this program and to avoid duplicit computations, the critical sections of this program should be when P writes elements to the queue, and when either C1 or C2 reads from the queue. This would work fine, but there is a problem.

Suppose that the program was implemented naively, making the critical section of P, C1, and C2 are quite lengthy. During execution, P locks the mutex, produces data, writes to the queue, and releases the mutex. Then C1 gets to execute, going in to its critical section. While C1 is in its critical section, C2 gets scheduled to execute. Due to C1 still being in the critical section, C2 cannot get the lock, and thus cannot execute. A bit later, C1 finishes the execution of its critical section, and unlocks the mutex. Then P executes and goes into its critical section. While P is in its critical section, C2 gets scheduled to execute. Since P is in its critical section, C2 cannot get the lock and cannot execute. Then P finishes execution in its critical section, and unlocks the mutex. Then C1 goes next, and the cycle repeats. We see that C2 is never able to do anything, due to the fact that either P or C1 has the lock when C2 tries to get it. This situation is known as starvation. Since that is not desirable, conditional variables and semaphores should be used to avoid starvation.

3.1 Waiting on the conditional variable to change

Conditional variables are used to ensure that threads wait until a specific condition occurs. Using the example presented in the previous section, answer the following questions:

- What is the minimum number of conditions needed for the example to work as intended?
- What would those conditions be, and which thread(producer or consumer) should wait on that condition?

To use conditional variables, the following function calls are needed:

```
#include <pthread.h>
```

```
int test_var;
pthread_cond_t generic_condition;
pthread_mutex_t lock;

void *genericThread0(void *args){
  pthread_mutex_lock(&lock);
  //do awesome stuff
  pthread_cond_signal(&generic_condition);
  test_var = 1;
  pthread_mutex_unlock(&lock);
}

void *genericThread1(void *args){
  pthread_mutex_lock(&lock);
  while(test_var == 0){
    pthread_cond_wait(&generic_condition, &lock);
  }
  //does fun things
  pthread_mutex_unlock(&lock);
}
.
.
.
int main(int argc, char **argv){
  test_var = 0;
.
  err = pthread_mutex_init(&lock, NULL);
.
.
  err = pthread_cond_init(&generic_condition, NULL);
.
.
.
 err = pthread_cond_destroy(&generic_condition);
 return 0;
}
```

For more information regarding the cond_init, cond_wait, cond_signal(and in extension cond_broadcast), and cond_destroy please consult `man 3 pthread_cond_init`, `man 3 pthread_cond_wait`, `man 3 pthread_cond_signal`, and `man 3 pthread_cond_destroy` respectively.

As can be seen above, the variable `generic_condition` is declared as a global variable for the same reason that `lock` is declared as a global variable. `generic_condition` is then initialized in main by calling `pthread_cond_init`. `genericThread0` locks the mutex, does what it's supposed to do, then sets `test_var` to `1`, so that `genericThread1` can break out of the loop, signals the conditional variable, then unlocks the mutex.

`genericThread1` will attempt to lock the mutex, test the value of `test_var`, and call `pthread_cond_wait` to see if the conditional variable has been signaled. If not, the thread will block, and `pthread_cond_wait` will not return. However, according to the man pages, this block does not last forever, and should be re-evaluated each time `pthread_cond_wait` returns; hence the `while` loop that surrounds the call to `pthread_cond_wait`. If the conditional variable has been signaled, then `pthread_cond_wait` would return, and the thread calling it would get the mutex. The value of `test_var` would then be tested, fall through, fun things are performed, and the mutex is unlocked.

Once all is said and done, remove the conditional variable using `pthread_cond_destroy`.

To see an example of conditional variables, please take a look at `cond_example.c`. Make sure that everything pertaining to condition variables, such as how it's created, and used, is understood before compiling it. Run the compiled code, and put the output value of the program into your report.

3.2 Why not semaphore; you've got to declare yourself openly

Semaphores perform a similar task to conditional variables, and they are slightly easier to use. Semaphores come it two flavors, Named, and Unnamed. The differences between the two are in how they are created, and destroyed. For simplicity, the unnamed flavor of semaphores will be covered in this handout. To use semaphores, the following function calls and includes are needed:

```
#include <semaphore.h>

sem_t semaphore;
.
.
.
void *genericThread0(void *args){
 pthread_mutex_lock(&lock);
 err = sem_wait(&semaphore);
 ...
 //do awesome stuff
 err = sem_post(&semaphore);
 ...
 pthread_mutex_unlock(&lock);
}

void *genericThread1(void *args){
 err = sem_wait(&semaphore);
 pthread_mutex_lock(&lock);
 ...
 //does fun stuff
 err = sem_post(&semaphore);
 ...
 pthread_mutex_unlock(&lock);
}

int main(int argc, char **argv){
 .
 .
 .
 err = sem_init(&semaphore, 0, 1);
 .
 .
 .
 err = sem_destroy(&semaphore);
 ...
 return 0;
}
```

For more information on the init, wait, post, and destroy functions, consult `man 3 sem_init`, `man 3 sem_wait`, `man 3 sem_post`, and `man 3 sem_destroy` respectively.

Note that the calls to`pthread_mutex_lock` and `pthread_mutex_unlock` do not necessarily have to be where they are shown in the above code, i.e., the mutex lock does not have to occur before the semaphore wait, and the mutex unlock doesn't have to occur after the semaphore post.

For similar reasons to `lock` and `generic_condition`, `semaphore` is declared as a global variable. It is initialized in main using `sem_init` with the value for *pshared* set to `0`(meaning the semaphore is only shared between the threads of the current process), and it's initial value will be `1`(last argument to `sem_init`).

`genericThread0` decrements `semaphore` by one using `sem_wait`. If `genericThread1` attempts to decrement `semaphore` when it got to its call to `sem_wait` before `genericThread0` incremented the semaphore by calling `sem_post`, then `genericThread1` will block right at its `sem_wait` line. This is because decrementing a semaphore past `0` will not evaluate. So if a semaphore is already at a value of `0`, decrementing it with `sem_wait` will cause the thread calling `sem_wait` to wait until the value of the semaphore is incremented to a value greater than `0`.

Once everything is completed, destroy the semaphore by calling `sem_destroy` on `semaphore`.

For an example of semaphores being used, take a look at `sem_example.c`. Please make sure that everything in the program `sem_example.c` pertaining to semaphores is understood before compiling it. Run the program, and note the order in which the buffer is read/written to. Run the program multiple times, and again not the order in which the buffer is read/written to. Do they look different? Why do you think that is the case?

4 Tasks for this lab

This is week one of a three part lab assignment. You will be expected to submit this week's work as normal with a lab report. Next week's lab will build on the code you have written for this lab. It is therefore in your best interest to write clean and portable code to minimize work in future labs.

4.1 Print Server

This week's lab is to write a multi-threaded print server program. The server will take in jobs from stdin in a Postscript format and print the files to PDF. The layout of the program is one producer thread reading files from stdin and inserting the jobs into one of two queues. Each queue has a set of consumer threads which will take the print jobs and print to PDF. Since the printing takes time we will want to print multiple files at the same time; hence we are using multiple threads for this problem. The overall flow is shown below:

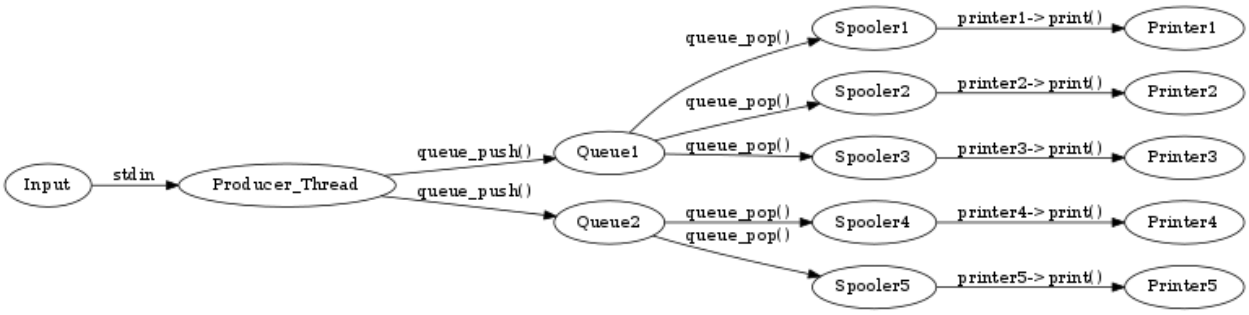

Figure 1:

The idea being simulated here is a single print server printing to a number of printers. Each print job will request what type of printer it should be print on (e.g. color vs. black and white). There may be multiple of a single type of printer; e.g. there could be three BW printers and one color, and a BW job could be printed to any BW printer. For this first lab there is really only one type of printer (ps2pdf), but we are going to install this driver twice to simulate two different types of printers. In a future lab we will expand this to

additional types of printers. To implement this define two print queues, one for each type of printer (pdf1, and pdf2) and install the pdf printer driver n1+n2 times, but allocate n1 to the pdf1 queue and n2 to the pdf2 queue. This is explained in the code as well.

4.2 Given Code

For this lab a large portion of the code is already written and supplied to you. You will first need to read through the code provided in the `./src/` directory to understand what is already finished and what needs to be added. You will also find, as is often the case when you are given code, that you will need to fix some of the given code to meet the project expectations. For example, you will find that some of the errors are reported in a way that will not meet the –quiet command line flag you must support. You should NOT change the existing API (i.e. the function prototypes or structure types) unless a comment tells you to do so. If you do change you sould only add to maintain backwards compatability. To help understand the code you can look at the doxygen generated comments. To generate the documentation run `make doc` from inside the `src` directory. The documentation can also be found at http://cpre308.github.io/labs/Lab5/doc/html/index.html.

Below is a brief description of each file. You still should look through all of the code to make sure you understand what it is doing.

4.2.1 main.c

This file is where the main function is along with the `printer_spooler` function. `main` will act as the producer thread reading jobs from stdin and sending them to the print queues. The `printer_spooler` function will be the consumer threads. Each `printer_spooler` will be assigned a single printer to print to (created by calling `printer_install()`) and a print job queue. The thead will be in an infinite loop pulling jobs off the queue and sending them to the printer.

4.2.2 queue.c, queue.h

These two files are used for representing queues. In this system a queue is a doubly linked list with functions to push and pop elements on both the head and tail of the list. The student will need to fill in all of the `todo` sections of these files to add mutual exclusion to the list. Look at the files for more information. *Note*, currently the setup is to use semaphores to implement this functionallity. If you would like you can change this to using condition variables instead, or you can stick with the semaphores.

4.2.3 print_job.c, print_job.h

These two files represent a print job. The function `print_job_create` creates a new print job object and allocates a temporary file in the `/tmp` directory. The student is expected to fill in the `print_job_tostring` function, and should read through the rest to understand what is going on.

4.2.4 printer.h

This file is the the public API of a printer driver. Any print driver that is later used with this server will need to include this header. In the next labs the student will write an additional print driver and load that driver into this program at runtime. For now just pass `NULL` into the first param of `printer_install`.

4.2.5 pdf_printer.c

This file is an example print driver for this print server. For this week the student will just compile the driver into the program. In a future lab this will be compiled seperately and loaded at runtime. This is why the `print` and `uninstall` methods are accessed through the `printer_t` object.

4.2.6 Makefile

Everything you have done for this class so far were simple one file programs. This time the project is larger and being split into multiple files. The supplied makefile will compile all of the source files and link them together to create a executable file. You should look at the makefile, but you do not have to make any changes to it for this lab. This make file can do the following:

```
# compile the code
$ make
# remove all binaries and object files
$ make clean
# generate the documentation
$ make doc
# make a single object file
$ make queue.o
```

4.3 Input file format

The files to be printed will be supplied through stdin. The streams will begin with several lines of header information, each starting with a `#` sign. After the last header information the actual Postscript file will begin with at `%` sign. The end of the Postscript file will end with a `%EOF`. See the example files to understand the format, and the `main.c` file to find more information on how to parse the file.

4.4 What you should do

First read all of the code and understand what it currently does and read the rest of this document. Go through the code and fix all of the `todo` tags and implement the functionality of the program. The provided test script only tests a very basic case; you should write a better test script that will test all of the features of the code. You should write a report answering all questions in this lab write up and detailing how you solved this problem. If you made any changes to the code explain what you changed and why. Talk about any issues you ran into and how you overcame them.

4.5 What to submit

Submit the following via GitHub:

- Any code you have written.
  - Your code should be well commented with doxygen style comment
  - Your code must be backwards compatible with the supplied API
- The test script you wrote to test and any additional files it takes (such as additional .ps files).
- Your lab report
- A README explaining how to use your test script

4.6 Additional resources

The man pages are going to be your best friend for this lab. In addition to the man pages, the following sites will also be helpful:

- https://computing.llnl.gov/tutorials/pthreads/
- https://www.gnu.org/software/libc/manual/pdf/libc.pdf
- http://cpre308.github.io/labs/Lab5/doc/html/index.html
- https://gcc.gnu.org/onlinedocs/gcc-4.9.2/gcc.pdf

4.7 Extra credit

There are a couple of ways to get extra credit for this lab. One is to support additional printer drivers. For example you could use ghostwriter to take in a Postscript file and output an ascii text file (`ps2ascii`). Another way to get extra credit is to support additional **useful** command line arguments. Please document any additional features in your lab report so the TAs know to grade them.

5 License

This lab write up and all accompany materials are distributed under the MIT License. For more information, read the accompaning LICENSE file distributed with the source code.