CS423: Operating Systems Design
MP1: Introduction to Linux Kernel Programming

1 Goals and Overview

- In this Machine problem you will learn the basics of Linux Kernel Programming
- You will learn to create a Linux Kernel Module (LKM).
- You will use Timers in the Linux Kernel to schedule work
- You will use Workqueues to defer work through the use of the Two-Halves concept
- You will use the basics of the Linked Lists interface in the Linux Kernel to temporarily store data in the kernel
- You will learn the basic of Concurrency and Locking in the Linux Kernel
- You will interface applications in the user space with your Kernel Module through the Proc Filesystem

2 Introduction

Kernel Programming has some particularities that can make it more difficult to debug and learn. In this section we will discuss a few of them.

The most important difference between kernel programming in Linux and Application programming in the user space is the **lack of Memory Protection**. That is driver, modules, and kernel threads all share the same memory address space. De-referencing a pointer that contains the wrong memory location and writing to it can cause the whole system to crash or corrupt important subsystems including filesystem and networking.

Another important difference is that in kernel programming, **preemption is not always available**, that means that we can indefinitely hog the CPU or cause system-wide deadlocks. This makes concurrency much more difficult to handle in the kernel,

than in user space. For example in kernel space we are responsible for ensuring that interrupt handlers are as efficient as possible using the CPU for very little time.

Also another important issue is the **lack of user space libraries**. Glib, C++ Standard Library and other libraries reside in the user space and cannot be accessed in the kernel. This limits what we can do and how do we implement it. Another important difference is that Linux Kernel lacks from Floating-Point support. That is all the math must be implemented using integers. Also files, signals or security descriptors are not available.

Through the rest of the document and your implementation you will learn some of the basic mechanisms, structures and designs common to many areas of Linux Kernel Development. **Please consult the recommended links and tutorials in the References Section of this document as those documents detail everything that you need to implement this MP.**

3 Developmental Setup

In this assignment, you will again work on the provided Virtual Machine and you will develop kernel modules for the Linux Kernel (with your *NetId* in the name) that you installed on your machine as part of MP0. Again, you will have full access and control of your Virtual Machine, you will be able to turn it on, and off using the VMWare vSphere Console. Inside your Virtual Machine you are free to install any editor or software like Gedit, Emacs and Vim.

During the completion of this and other MPs, errors in your kernel modules could potentially lead to "bricking" your VM, rendering it unusable. If this happens, please post to Piazza asking for help and a TA will work with Engr-IT to have your VM restored. However, this will cost you precious hours of development time! Fortunately, these problems can be almost entirely avoided by taking the following precautions:

- **VM Snapshots:** Before installing kernel module, be sure to use the vSphere console to take a snapshot of your VM. If the VM becomes unusable, you can go back to vSphere and restore from this snapshot. *A note on snapshots – snapshots are NOT backups, and can go to unbounded sizes as you continue to use your VM. Use snapshots sparingly and delete all of your snapshots at the conclusion of each assignment.*

- **Code Versioning:** If your VM becomes unusable, it is possible that Engr-IT will need to reimage your machine, causing you to lose any work stored on it.

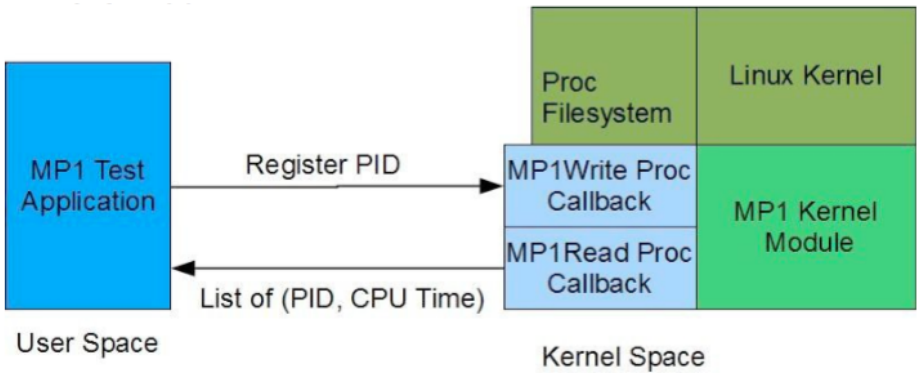

Figure 1: Proc Filesystem Interface between Test Application and MP1 Kernel Module

It is strongly recommended that you prevent loss of work by using a **private** `git` repo for your MP code, which can be easily setup on `github.com` or `bitbucket.org`.

- **Don't DoS /var/log:** During the course of our MPs we will be generating a *lot* of log messages, which eventually find their way to the `/var/log` directory on our disk. The `/var` partition on our VM contains about 4GB of space, and if it fills up then certain applications (e.g., `ssh`) may stop working. Keep an eye on the `/var` partition using the `df -h` command and remove large log files if it starts getting too full. Another workaround is to symlink /var/log to your main partition as follows:

  ```
  mkdir /srv/log
  rsync -auv /var/log/ /srv/log/
  rm -rf /var/log
  ln -s /srv/log /var/log
  reboot
  ```

  This will increase the space available to `/var/log`, but it is still possible to fill up the available space if you get caught in an infinite logging loop, etc., so keep an eye on this regardless.

4 Problem Description

In this MP you will build a kernel module that measures the User Space CPU Time of processes registered within the kernel module and a simple test case application that requests this service. In a real scenario many applications might be using this functionality implemented by our new kernel module and therefore our module is designed to support multiple applications/processes to register simultaneously.

The kernel module will allow processes to register themselves through the Proc Filesystem. For each registered process, the kernel module should write to an entry in the Proc Filesystem, the application's User Space CPU Time (known also as user time). The kernel module must keep these values in memory for each registered process and update them every 5 seconds. Figure 1 shows the application interface with the kernel module using the Proc filesystem.

The registration process must be implemented as follows: At the initialization of your kernel module, it must create a directory entry within the Proc filesystem (e.g. `/proc/mp1`). Inside this directory your kernel module must create a file entry (e.g. `/proc/mp1/status`), readable and writable by anyone (mask 0666). Upon start of a process (e.g. our test application), it will register itself by writing its PID to this entry that you created. When a process reads from this entry, the kernel module must print a list of all the registered PIDs in the system and its corresponding User Space CPU Times. An example of the format your proc filesystem entry can use to print this list is as follows:

`PID1`: *CPU Time of PID1*
`PID2`: *CPU Time of PID2*

Your kernel module implementation must store the PIDs and the CPU Time values of each process in a Linked List using the implementation provided by the Linux kernel. Part of the goals of this MP is that you learn to use this facility provided by the kernel. Additionally, the CPU Time values of each process must be periodically updated by using a Kernel Timer. However, you must use a technique called **Two-Halves** approach. In this approach an interrupt is divided in two parts: Interrupt Handler (Top-Half) and Thread performing the work (Bottom-Half). The slides on Linux Kernel Programming, posted on compass explain this concept more in detail.

In our case the Top-Half will be the Timer Interrupt Handler, its sole purpose will be to wake up the Bottom-Half. For the Bottom-Half we will use a work function in a workqueue. A workqueue is a kernel mechanism that allows you to schedule the

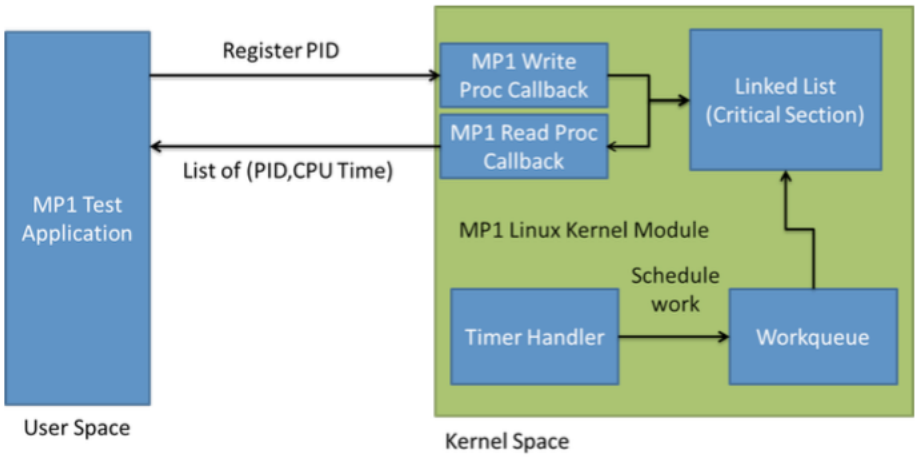

Figure 2: MP1 Architecture Overview

execution of a function (work function) at a later time. A worker thread managed by the kernel is responsible of the execution of each of the work functions scheduled in the workqueue. In our MP, the work function will traverse the link list and update the CPU Time values of each registered process.

It is acceptable for your work function to be scheduled even if there are no registered processes, however you might consider an implementation where the timer is not scheduled if there are no registered processes. Figure 2 shows the architecture of the kernel module you should implement, including the timer interrupt and the workqueue.

Finally, the kernel module will be responsible of freeing any resource that it allocated during its execution. This includes stopping any work function pending, destroying any timer, workqueue or Proc filesystem entry created and freeing any allocated memory.

As a test case you must implement a simple program that registers itself in the kernel module using the proc filesystem and then calculates a series of factorial computations. This computation can repeat or can be different. However, this program should run for sufficient time to test your kernel module, sometime between 10 and 15 seconds should be sufficient. At the end of the computation the application must read the proc file system entry containing the list of all the registered applications and its corresponding CPU Times.

5 Implementation Challenges

In this MP you will find many challenges commonly found in Kernel Programming. Some of these challenges are discussed below:

- During the registration process you will need to **access data from the User Space**. Kernel and Applications both run in **two separate memory spaces**, so de-referencing pointers containing data from the User Space is not possible. Instead you must use the function copy_from_user() to copy the data into a buffer located in kernel memory. Similarly, when returning data through a pointer we must copy the data from the kernel space into the user space using the function copy_to_user(). Common cases where this might appear are in Proc filesystem callbacks and system calls.
- Another important challenge is the **lack of libraries**, instead the kernel provides similar versions of this commonly used functions found in libraries. For example malloc() is replaced with kmalloc(), printf() is replaced by printk(). Some other handy functions implemented in the kernel are sprintf(), and sscanf(). This functions are introduced in [2].
- Throughout your implementation, you will need to face **different running contexts**. A context is the entity whom the kernel is running code on behalf of. In the Linux kernel you will find 3 different contexts:
    1. Kernel Context: Runs on behalf of the kernel itself. Example: Kernel Threads and workqueues
    2. ProcessContext: Runs on behalf of some process. Example: SystemCalls
    3. InterruptContext: Runs on behalf of an interrupt. Example: TimerInterrupt
- The Linux kernel is a **preemptible kernel**. This means that all the contexts run concurrently and can be interrupted from its execution at any time. You will need to protect your data structures through the use of appropriate **locks** and prevent race conditions wherever they appear. Please note that architectural reasons limit which type of locks can be used for each context. For example, **interrupt context code cannot sleep and therefore semaphores will create a deadlock when used in this case.**

This sleeping restriction in interrupt context also prevents you from using various functions that sleep during its execution. Many of these functions involve complicated operations that require access to devices like printk(), functions that schedule processes, copy data from and to the user space, and functions that allocate memory (e.g

kmalloc()). Some exceptions to this rule of thumb are the function wake_up_process() and the function kmalloc() when used with special flags.

Due to all these challenges, we recommend you that you test your code often and build in small increments. You can use the BUG_ON() macro to spot inconsistencies and trigger a stack dump.

6 Implementation Overview

In this section we will briefly guide you through the implementation. Figure 2 shows the architecture of MP1, showing the kernel module with its workqueue and timer and also the proc filesystem all in the kernel space. In the user space you can see the test application that you will also implement.

Step 1: The best way to start is by implementing an **empty ('Hello World!') Linux Kernel Module**.

Step 2: After this you should implement the **Proc Filesystem entries** (i.e /proc/mp1/ and /proc/mp1/status). Make sure that you implement the creation of these entries in your module "init" function and the destruction in your module ?exit? function.

At this point you should probably test your code. Compile the module and load it in memory using insmod or modprobe. You should be able to see the proc filesystem entries you created using ls. Now remove the module and check that the entries are properly removed.

Step 3: The next step should be to implement the **full registration**, you will need to declare and initialize a Linux kernel Linked List. The kernel provides macros and functions to traverse the list, and insert and delete elements.

Step 4: You will also need to implement the **callback functions for read and write** in the entry of the proc filesystem you created. Keep the format of the registration string simple. We suggest that a user space application should be able to register itself by simply writing the PID to the proc filesystem entry you created (e.g ?pid?>/mp1/status). The callback functions will read and write data from and to the user space so you need to use copy_from_user() and copy_to_user(). To keep things simple, do not worry about adding support for page breaks in the reading callback.

Step 5: At this point you should be able to write a **simple user space application that registers itself** in the module. Your test application can use the function getpid() to obtain its PID. You can open and write to the proc filesystem entry using fopen() and fprintf(), or you can use sprintf() and the system() function to execute the string echo ?pid?>/mp1/status in the command line.

Step 6: The next step should be to create a **Linux Kernel Timer** that wakes up every 5 seconds. Timers in the kernel are single shot (i.e not periodic). Expiration times for Timers in Linux are expressed in **jiffies** and they refer to an absolute time since boot. Jiffy is a unit of time that expresses the number of clock ticks of the system timer in Linux. The conversion between seconds and jiffies is system dependent and can be done using the constant **Hz**. The global variable jiffies can be used to retrieve the current time elapsed since boot expressed in jiffies.

Step 7: Next you will need to implement the **work function**. At the timer expiration, the timer handler must use the workqueue API to schedule the work function to be executed as soon as possible. To test your code you can use printk() to print to the console every time the work function is executed by the workqueue worker thread. You can see these messages by using the command dmesg in the command line. Also please note that the workqueue API was updated for kernel 2.6.20 and newer, therefore some documentation about workqueues on the internet might be outdated.

Step 8: Now, you will need to implement the **updates to the CPU Times** for the processes in the Linked List. We have provided in the file mp1_given.h a helper function int get_cpu_use(int pid, unsigned long* cpu_value) to simplify this part. This function returns 0 if the value was successfully obtained and returned through the parameter cpu_value, otherwise it returns -1. As part of the update process, you will need to use locks to protect the Linked List and any other shared variables accessed by the three contexts (kernel, process, interrupt context). The advantage of using a two half approach is that in most cases the locking will be placed in the **work function** and not in the timer interrupt. If a registered process terminates, get_cpu_use will return -1. In this case, the registered process should be removed from the linked list.

Step 9: Finally you should check for **memory leaks** and make sure that everything is **properly deallocated** before we exit the module. Please keep in mind that need to stop any asynchronous entity running (e.g timers, thread, workqueues) before deallocating memory structures. At this time, kernel module coding is finished. Now you should be able to **implement the factorial test application** and have some additional testing of your code.

7 Software Engineering

Your code should include comments where appropriate. It is not a good idea to repeat what the function does using pseudo-code, but instead, provide a high-level overview of the function including any preconditions and post-conditions of the algorithm. Some

functions might have as few as one line comments, while some others might have a longer paragraph.

Also, your code must be split into small functions, even if these functions contain no parameters. This is a common situation in kernel modules because most of the variables are declared as global, including but not limited to data structures, state variables, locks, timers and threads.

An important problem in kernel code readability is to know if a function holds the lock for a data structure or not, different conventions are usually used. A common convention is to start the function with the character '_' if the function does not hold the lock of a data structure.

In kernel coding, performance is a very important issue; usually the code uses macros and preprocessor commands extensively. Proper use of macros and identifying possible situations where they should be used is important in kernel programming.

Finally, in kernel programming, the use of the goto statement is a common practice. A good example of this, is the implementation of the Linux scheduler function schedule(). In this case, the use of the goto statement improves readability and/or performance. "Spaghetti code" is never a good practice.

8 Submission Instructions

Here are the steps required to submit your MP. *Failure to correctly follow these instructions will result in a deduction from your grade (-10pts). No late submissions are accepted!*

0) At the time this assignment is due, your VM needs to be up and running in your custom kernel so that we can SSH into it and run your code. Do not use the VM until the instructor informs you that it is safe to use it again after the conclusion of grading.

1) In your home directory, create a directory `MP1`. In this directory, copy all of your source code, a Makefile that compiles your code into a kernel module named **NETID_MP1.ko** (replace ID with your own NetID), a `README` file that includes a complete description of how to run your program, and a screenshot of the output when running the command "cat /proc/mp1/status". At least two of your user program instances need to be running when capturing the screenshot. To run two user program instances concurrently, do "./userapp & ./userapp &" in your terminal (assume your user program is named userapp).

2) From your home directory, create a tarball named **NETID_MP1.tar.gz** by running the command `tar czvf NETID_MP1.tar.gz MP1`. Again, replace NETID here with your own NetID. Then, as root create a directory `MPs` in the top directory of root partition, then copy **NETID_MP1.tar.gz** into this directory. The full path to your submission should be `/MPs/NETID_MP1.tar.gz`. When we unzip the tarball, we should find your code and a Makefile in `/MPs/MP1/`.

3) Copy the **NETID_MP1.tar.gz** file you created off of your machine and submit it on Compass2g. Make sure that your Compass2g submission and the tarball at `/MPs/NETID_MP1.tar.gz` are *identical*, as we will be comparing the hashes of these two files.

9 Grading Criteria

Criterion	**Points**
Can we insert your module?	5
Does your user app function correctly?	5
Does proc read work correctly?	15
Does proc write work correctly?	15
Does the interrupt handler work correctly (incl. removing finished processes)?	15
Does your module correctly support multiple processes?	10
Is your critical region lock correctly implemented?	5
Does your module correctly free all memory?	10
Can we remove your module (rmmod)?	5
Documented code and README file	5
Your code compiles and runs correctly and does not use any Floating Point arithmetic.	5
Your code is well commented, readable and follows software engineering principles.	5
Total	**100**

10 Deadline

Refer to Compass2G for the official deadline for this assignment.

11 References

1. The Kernel Newbie Corner: Kernel Debugging Using proc "Sequence" Files
   https://www.linux.com/learn/linux-training/37985-the-kernel-newbie-corner-kernel-debugging-using-proc-qsequenceq-files-part-1
   http://www.linux.com/learn/linux-career-center/39972-kernel-debugging-with-proc-qsequenceq-files-part-2-of-3
   http://www.linux.com/learn/linux-career-center/44184-the-kernel-newbie-corner-kernel- debugging-with-proc-qsequenceq-files-part-3

2. The Linux Kernel Module Programming Guide (a little outdated, but still useful)
   http://tldp.org/LDP/lkmpg/2.6/html/index.html

3. Linux Kernel Linked List Explained
   http://isis.poly.edu/kulesh/stuff/src/klist/

4. Kernel API's Part 3: Timers and lists in the 2.6 kernel
   http://www.ibm.com/developerworks/linux/library/l-timers-list/

5. Access the Linux Kernel using the Proc Filesystem, (a little outdated, still useful)
   http://www.ibm.com/developerworks/linux/library/l-proc/index.html

6. Kernel APIs Part2: Deferrable functions, kernel tasklets, and work queues
   http://www.ibm.com/developerworks/linux/library/l-tasklets/index.html

7. Love Robert, Linux Kernel Development, Chapters 6, 8-11, 17-18, Addison-Wesley Professional, Third Edition

8. Linux synchronization methods (2.6 kernel)
   http://www.makelinux.net/ldd3/chp-5-sect-3
   http://www.makelinux.net/ldd3/chp-5-sect-5