Lectures on Operating Systems (Mythili Vutukuru, IIT Bombay)

Practice Problems: Concurrency

1. Answer yes/no, and provide a brief explanation.

   (a) Is it necessary for threads in a process to have separate stacks?

   (b) Is it necessary for threads in a process to have separate copies of the program executable?

   Ans:

   (a) Yes, so that they can have separate execution state, and run independently.

   (b) No, threads share the program executable and data.

2. Can one have concurrent execution of threads/processes without having parallelism? If yes, describe how. If not, explain why not.

   Ans:

   Yes, by time-sharing the CPU between threads on a single core.

3. Consider a multithreaded webserver running on a machine with N parallel CPU cores. The server has M worker threads. Every incoming request is put in a request queue, and served by one of the free worker threads. The server is fully saturated and has a certain throughput at saturation. Under which circumstances will increasing M lead to an increase in the saturation throughput of the server?

   Ans: When $M < N$ and the workload to the server is CPU-bound.

4. Consider a process that uses a user level threading library to spawn 10 user level threads. The library maps these 10 threads on to 2 kernel threads. The process is executing on a 8-core system. What is the maximum number of threads of a process that can be executing in parallel?

   Ans: 2

5. Consider a user level threading library that multiplexes $N > 1$ user level threads over $M \geq 1$ kernel threads. The library manages the concurrent scheduling of the multiple user threads that map to the same kernel thread internally, and the programmer using the library has no visibility or control on this scheduling or on the mapping between user threads and kernel threads. The N user level threads all access and update a shared data structure. When (or, under what conditions) should the user level threads use mutexes to guarantee the consistency of the shared data structure?

   (a) Only if $M > 1$.

   (b) Only if $N \geq M$.

   (c) Only if the M kernel threads can run in parallel on a multi-core machine.

(d) User level threads should always use mutexes to protect shared data.

Ans: (d) (because user level threads can execute concurrently even on a single core)

6. Which of the following statements is/are true regarding user-level threads and kernel threads?

   (a) Every user level thread always maps to a separate schedulable entity at the kernel.

   (b) Multiple user level threads can be multiplexed on the same kernel thread

   (c) Pthreads library is used to create kernel threads that are scheduled independently.

   (d) Pthreads library only creates user threads that cannot be scheduled independently at the kernel scheduler.

   Ans: (b), (c)

7. Consider a Linux application with two threads T1 and T2 that both share and access a common variable x. Thread T1 uses a `pthread` mutex lock to protect its access to x. Now, if thread T2 tries to write to x without locking, then the Linux kernel generates a trap. [T/F]

   Ans: F

8. In a single processor system, the kernel can simply disable interrupts to safely access kernel data structures, and does not need to use any spin locks. [T/F]

   Ans: T

9. In the `pthread` condition variable API, a process calling wait on the condition variable must do so with a mutex held. State one problem that would occur if the API were to allow calls to wait without requiring a mutex to be held.

   Ans: Wakeup happening between checking for condition and sleeping causing missed wakeup.

10. Consider N threads in a process that share a global variable in the program. If one thread makes a change to the variable, is this change visible to other threads? (Yes/No)

    Ans: Yes

11. Consider N threads in a process. If one thread passes certain arguments to a function in the program, are these arguments visible to the other threads? (Yes/No)

    Ans: No

12. Consider a user program thread that has locked a pthread mutex lock (that blocks when waiting for lock to be released) in user space. In modern operating systems, can this thread be context switched out or interrupted while holding the lock? (Yes/No)

    Ans: Yes

13. Repeat the previous question when the thread holds a pthread spinlock in user space.

    Ans: Yes

14. Consider a process that has switched to kernel mode and has acquired a spinlock to modify a kernel data structure. In modern operating systems, will this process be interrupted by external hardware before it releases the spinlock? (Yes/No)

    Ans: No

15. Consider a process that has switched to kernel mode and has acquired a spinlock to modify a kernel data structure. In modern operating systems, will this process initiate a disk read before it releases the spinlock? (Yes/No)

    Ans: No

16. When a user space process executes the wakeup/signal system call on a pthread condition variable, does it always lead to an immediate context switch of the process that calls signal (immediately after the signal instruction)? (Yes/No)

    Ans: No

17. Consider a process in kernel mode that acquires a spinlock. For correct operation, it must disable interrupts on its CPU core for the duration that the spinlock is held, in both single core and multicore systems. [T/F]

    Ans. T

18. Consider a process in kernel mode that acquires a spinlock in a multicore system. For correct operation, we must ensure that no other kernel-mode process running in parallel on another core will request the same spinlock. [T/F]

    Ans. F

19. Multiple threads of a program must use locks when accessing shared variables even when executing on a single core system. [T/F]

    Ans: T

20. Recall that the atomic instruction compare-and-swap (CAS) works as follows:
    `CAS(&var, oldval, newval)` writes `newval` into `var` and returns true if the old value of `var` is `oldval`. If the old value of `var` is not `oldval`, `CAS` returns false and does not change the value of the variable. Write code for the function to acquire a simple spinlock using the `CAS` instruction.

    Ans: while(!CAS(&lock, 0, 1));

21. The simple spinlock implementation studied in class does not guarantee any kind of fairness or FIFO order amongst the threads contending for the spin lock. A ticket lock is a spinlock implementation that guarantees a FIFO order of lock acquisition amongst the threads contending for the lock. Shown below is the code for the function to acquire a ticket lock. In this function, the variables `next_ticket` and `now_serving` are both global variables, shared across all threads, and initialized to 0. The variable `my_ticket` is a variable that is local to a particular thread, and is not shared across threads. The atomic instruction `fetch_and_increment(&var)` atomically adds 1 to the value of the variable and returns the old value of the variable.

    ```
    acquire():
      my_ticket = fetch_and_increment(&next_ticket)
      while(now_serving != my_ticket); //busy wait
    ```

    You are now required to write the code to release the spinlock, to be executed by the thread holding the lock. Your implementation of the release function must guarantee that the next contending

thread (in FIFO order) will be able to acquire the lock correctly. You must not declare or use any other variables.

```
release(): //your code here
```

Ans:

```
release(): //your code here
now_serving++;
```

22. Consider a multithreaded program, where threads need to aquire and hold multiple locks at a time. To avoid deadlocks, all threads are mandated to use the function `acquire_locks`, instead of acquiring locks independently. This function takes as arguments a variable sized array of pointers to locks (i.e., addresses of the lock structure), and the number of lock pointers in the array, as shown in the function prototype below. The function returns once all locks have been successfully acquired.

```
void acquire_locks(struct lock *la[], int n);
//i-th lock in array can be locked by calling lock(la[i])
```

Describe (in English, or in pseudocode) one way in which you would implement this function, while ensuring that no deadlocks happen during lock acquisition. Your solution must not use any other locks beyond those provided as input. Note that multiple threads can invoke this function concurrently, possibly with an overlapping set of locks, and the lock pointers can be stored in the array in any arbitrary order. You may assume that the locks in the array are unique, and there are no duplicates within the input array of locks.

Ans. Sort locks by address struct lock *, and acquire in sorted order.

23. Consider a process where multiple threads share a common Last-In-First-Out data structure. The data structure is a linked list of "struct node" elements, and a pointer "top" to the top element of the list is shared among all threads. To push an element onto the list, a thread dynamically allocates memory for the struct node on the heap, and pushes a pointer to this struct node in to the data structure as follows.

```
void push(struct node *n) {
n->next = top;
top = n;
}
```

A thread that wishes to pop an element from the data structure runs the following code.

```
struct node *pop(void) {
struct node *result = top;
if(result != NULL) top = result->next;
return result;
}
```

A programmer who wrote this code did not add any kind of locking when multiple threads concurrently access this data structure. As a result, when multiple threads try to push elements onto this structure concurrently, race conditions can occur and the results are not always what one would expect. Suppose two threads T1 and T2 try to push two nodes n1 and n2 respectively onto the data structure at the same time. If all went well, we would expect the top two elements of the data structure would be n1 and n2 in some order. However, this correct result is not guaranteed when a race condition occurs.

Describe how a race condition can occur when two threads simultaneously push two elements onto this data structure. Describe the exact interleaving of executions of T1 and T2 that causes the race condition, and illustrate with figures how the data structure would look like at various phases during the interleaved execution.

Ans: One possible race condition is as follows. n1's next is set to top, then n2's next is set to top. So both n1 and n2 are pointing to the old top. Then top is set to n1 by T1, and then top is set to n2 by T2. So, finally, top points to n2, and n2's next points to old top. But now, n1 is not accessible by traversing the list from top, and n1 remains on a side branch of the list.

24. Consider the following scenario. A town has a very popular restaurant. The restaurant can hold N diners. The number of people in the town who wish to eat at the restaurant, and are waiting outside its doors, is much larger than N. The restaurant runs its service in the following manner. Whenever it is ready for service, it opens its front door and waits for diners to come in. Once N diners enter, it closes its front door and proceeds to serve these diners. Once service finishes, the backdoor is opened and the diners are let out through the backdoor. Once all diners have exited, another batch of N diners is admitted again through the front door. This process continues indefinitey. The restaurant does not mind if the same diner is part of multiple batches.

    We model the diners and the restaurant as threads in a multithreaded program. The threads must be synchronized as follows. A diner cannot enter until the restaurant has opened its front door to let people in. The restaurant cannot start service until N diners have come in. The diners cannot exit until the back door is open. The restaurant cannot close the backdoor and prepare for the next batch until all the diners of the previous batch have left.

    Below is given unsynchronized pseudocode for the diner and restaurant threads. Your task is to complete the code such that the threads work as desired. Please write down the complete synchronized code of each thread in your solution.

    You are given the following variables (semaphores and initial values, integers) to use in your solution. The names of the variables must give you a clue about their possible usage. You must not use any other variable in your solution.

    ```
    sem (init to 0): entering_diners, exiting_diners, enter_done, exit_done
    sem (init to 1): mutex_enter, mutex_exit
    Integer counters (init to 0): count_enter, count_exit
    ```

    All changes to the counters and other variables must be done by you in your solution. None of the actions performed by the unsynchronized code below will modify any of the variables above.

    (a) Unsynchronized code for the restaurant thread is given below. Add suitable synchronization in your solution in between these actions of the restaurant.

    ```
    openFrontDoor()
    closeFrontDoor()
    serveFood()
    openBackDoor()
    closeBackDoor()
    ```

    (b) Unsynchronized code for the diner thread is given below. Add suitable synchronization in your solution around these actions of the diner.

    ```
    enterRestaurant()
    eat()
    exitRestaurant()
    ```

Ans: Correct code for restautant thread:

```
openFrontDoor()
do N times: up(entering_diners)
down(enter_done)

closeFrontDoor()
serveFood()

openBackDoor()
do N times: up(exiting_diners)
down(exit_done)
closeBackDoor()
```

Correct code for the diner thread:

```
down(entering_diners)
enterRestaurant()

down(mutex_enter)
  count_enter++
  if(count_enter == N) {
    up(enter_done)
    count_enter = 0
    }
up(mutex_enter)

eat()

down(exiting_diners)
exitRestaurant()

down(mutex_exit)
  count_exit++
  if(count_exit == N) {
    up(exit_done)
    count_exit = 0
    }
up(mutex_exit)
```

An alternate to doing up N times in restaurant thread is: restaurant does up once, and every woken up diner does up once until N diners are done. This alternate solution is shown below. Correct code for restautant thread:

```
openFrontDoor()
up(entering_diners)
down(enter_done)

closeFrontDoor()
serveFood()

openBackDoor()
up(exiting_diners)
down(exit_done)
closeBackDoor()
```

Correct code for the diner thread:

```
down(entering_diners)
enterRestaurant()

down(mutex_enter)
  count_enter++

  if(count_enter < N)
      up(entering_diners)
  else if(count_enter == N) {
    up(enter_done)
    count_enter = 0
    }
up(mutex_enter)

eat()

down(exiting_diners)
exitRestaurant()

down(mutex_exit)
  count_exit++
  if(count_exit < N)
     up(exiting_diners)
  else if(count_exit == N) {
    up(exit_done)
    count_exit = 0
    }
up(mutex_exit)
```

25. Consider a scenario where a bus picks up waiting passengers from a bus stop periodically. The bus has a capacity of K. The bus arrives at the bus stop, allows up to K waiting passengers (fewer if less than K are waiting) to board, and then departs. Passengers have to wait for the bus to arrive and then board it. Passengers who arrive at the bus stop after the bus has arrived should not be allowed to board, and should wait for the next time the bus arrives. The bus and passengers are represented by threads in a program. The passenger thread should call the function board() after the passenger has boarded and the bus should invoke depart() when it has boarded the desired number of passengers and is ready to depart.

    The threads share the following variables, none of which are implicitly updated by functions like board() or depart().

```
mutex = semaphore initialized to 1.
bus_arrived = semaphore initialized to 0.
passenger_boarded = semaphore initialized to 0.
waiting_count = integer initialized to 0.
```

    Below is given synchronized code for the passenger thread. You should not modify this in any way.

```
down(mutex)
waiting_count++
up(mutex)
down(bus_arrived)
board()
up(passenger_boarded)
```

    Write down the corresponding synchronized code for the bus thread that achieves the correct behavior specified above. The bus should board the correct number of passengers, based on its capacity and the number of those waiting. The bus should correctly board these passengers by calling up/down on the semaphores suitably. The bus code should also update waiting_count as required. Once boarding completes, the bus thread should call depart(). You can use any extra local variables in the code of the bus thread, like integers, loop indices and so on. However, you must not use any other extra synchronization primitives.

    Ans:

```
down(mutex)
N = min(waiting_count, K)
for i= 1 to N
  up(bus_arrived)
  down(passenger_boarded)
waiting_count = waiting_count - N
up(mutex)
depart()
```

26. Consider a roller coaster ride at an amusement park. The ride operator runs the ride only when there are exactly N riders on it. Multiple riders arrive at the ride and queue up at the entrance of the ride. The ride operator waits for N riders to accumulate, and may even take a nap as he waits. Once N riders have arrived, the riders call out to the operator indicating they are ready to go on the ride. The operator then opens the gate to the ride and signals exactly N riders to enter the ride. He then waits until these N riders enter the ride, and then proceeds to start the ride.

    We model the operator and riders as threads in a program. You must write pseudocode for the operator and rider threads to enable the behavior described above. Shown below is the skeleton code for the operator and rider threads. Complete the code to achieve the behavior described above. You can assume that the functions to open, start, and enter ride are implemented elsewhere, and these functions do what the names say they do. You must write the synchronization logic around these functions in order to invoke these functions at the appropriate times. You must use only locks and condition variables for synchronization in your solution. You may declare, initialize, and use other variables (counters etc.) as required in your solution.

```
//operator code, fill in the missing details

....
open_ride()
....
start_ride()
....

//rider thread, fill in the missing details
....
enter_ride()
....
```

Ans:

```
//variables: int rider_count (initialized to 0)
//variables: int enter_count (initialized to 0)
//condvar cv_rider, cv_operator1, cv_operator2
//mutex

//operator
lock(mutex)
while(rider_count < N) wait(cv_operator1, mutex)

open_ride()
do N times: signal(cv_rider)
while(enter_count < N) wait(cv_operator2, mutex)

start_ride()
unlock(mutex)

//rider
lock(mutex)
rider_count++
if(rider_count == N) signal(cv_operator1)
wait(cv_rider, mutex) // all wait, even N-th guy

enter_ride()

enter_count++
if(enter_count == N) signal(cv_operator2)
unlock(mutex)
```

27. A host of a party has invited $N > 2$ guests to his house. Due to fear of Covid-19 exposure, the host does not wish to open the door of his house multiple times to let guests in. Instead, he wishes that all N guests, even though they may arrive at different times to his door, wait for each other and enter the house all at once. The host and guests are represented by threads in a multi-threaded program. Given below is the pseudocode for the host thread, where the host waits for all guests to arrive, then calls openDoor(), and signals a condition variable once. You must write the corresponding code for the guest threads. The guests must wait for all N of them to arrive and for the host to open the door, and must call enterHouse() only after that. You must ensure that all N waiting guests enter the house after the door is opened. You must use only locks and condition variables for synchronization.

    The following variables are used in this solution: lock m, condition variables cv_host and cv_guest, and integer guest_count (initialized to 0). You must not use any other variables in the guest for synchronization.

```
//host
lock(m)
while(guest_count < N)
  wait(cv_host, m)
openDoor()
signal(cv_guest)
unlock(m)
```

    Ans:

```
//guest
lock(m)
guest_count++
if(guest_count == N)
  signal(cv_host)
wait(cv_guest, m)
signal(cv_guest)
unlock(m)
enterHouse()
```

28. Consider the classic readers-writers synchronization problem described below. Several processes/threads wish to read and write data shared between them. Some processes only want to read the shared data ("readers"), while others want to update the shared data as well ("writers"). Multiple readers may concurrently access the data safely, without any correctness issues. However, a writer must not access the data concurrently with anyone else, either a reader or a writer. While it is possible for each reader and writer to acquire a regular mutex and operate in perfect mutual exclusion, such a solution will be missing out on the benefits of allowing multiple readers to read at the same time without waiting for other readers to finish. Therefore, we wish to have special kind of locks called reader-writer locks that can be acquired by processes/threads in such situations. These locks have separate lock/unlock functions, depending on whether the thread asking for a lock is a reader or writer. If one reader asks for a lock while another reader already has it, the second reader will also be granted a read lock (unlike in the case of a regular mutex), thus encouraging more concurrency in the application.

    Write down pseudocode to implement the functions readLock, readUnlock, writeLock, and write-Unlock that are invoked by the readers and writers to realize reader-writer locks. You must use condition variables and mutexes only in your solution.

    Ans: A boolean variable `writer_present`, and two condition variables, `reader_can_enter` and `writer_can_enter`, are used.

```
readLock:
lock(mutex)
while(writer_present)
  wait(reader_can_enter)
read_count++
unlock(mutex)

readUnlock:
lock(mutex)
read_count--
if(read_count==0)
  signal(writer_can_enter)
unlock(mutex)

writeLock:
lock(mutex)
while(read_count > 0 || writer_present)
  wait(writer_can_enter)
writer_present = true
unlock(mutex)

writeUnlock:
lock(mutex)
writer_present = false
signal(writer_can_enter)
signal_broadcast(reader_can_enter)
unlock(mutex)
```

29. Consider the readers and writers problem discussed above. Recall that multiple readers can be allowed to read concurrently, while only one writer at a time can access the critical section. Write down pseudocode to implement the functions readLock, readUnlock, writeLock, and writeUnlock that are invoked by the readers and writers to realize read/write locks. You must use **only** semaphores, and no other synchronization mechanism, in your solution. Further, you must avoid using more semaphores than is necessary. Clearly list all the variables (semaphores, and any other flags/counters you may need) and their initial values at the start of your solution. Use the notation `down(x)` and `up(x)` to invoke atomic down and up operations on a semaphore `x` that are available via the OS API. Use sensible names for your variables.

Ans:

```
sem lock = 1; sem writer_can_enter = 1; int readCount = 0;

readLock:
down(lock)
readCount++
if(readCount ==1)
  down(writer_can_enter) //don't coexist with a writer
up(lock)

readUnlock:
down(lock)
readCount--
if(readCount == 0)
  up(writer_can_enter)
up(lock)

writeLock:
down(writer_can_enter)

writeUnlock:
up(writer_can_enter)
```

30. Consider the readers and writers problem as discussed above. We wish to implement synchronization between readers and writers, while giving **preference to writers**, where no waiting writer should be kept waiting for longer than necessary. For example, suppose reader process R1 is actively reading. And a writer process W1 and reader process R2 arrive while R1 is reading. While it might be fine to allow R2 in, this could prolong the waiting time of W1 beyond the absolute minimum of waiting until R1 finishes. Therefore, if we want writer preference, R2 should not be allowed before W1. Your goal is to write down pseudocode for read lock, read unlock, write lock, and write unlock functions that the processes should call, in order to realize read/write locks with writer preference. You must use only simple locks/mutexes and conditional variables in your solution. Please pick sensible names for your variables so that your solution is readable.

    Ans:

```
readLock:
lock(mutex)
while(writer_present || writers_waiting > 0)
    wait(reader_can_enter,mutex)
readcount++
unlock(mutex)

readUnlock:
lock(mutex)
readcount--
if(readcount==0)
    signal(writer_can_enter)
unlock(mutex)

writeLock:
lock(mutex)
writer_waiting++
while(readcount > 0 || writer_present)
    wait(writer_can_enter, mutex)
writer_waiting--
writer_present = true
unlock(mutex)

writeUnlock:
lock(mutex)
writer_present = false
if(writer_waiting==0)
    signal_broadcast(reader_can_enter)
else
    signal(writer_can_enter)
unlock(mutex)
```

31. Write a solution to the readers-writers problem with preference to writers discussed above, but using only semaphores.

Ans:

```
sem rlock = 1; sem wlock = 1;
sem reader_can_try = 1; sem writer_can_enter = 1;
int readCount = 0; int writeCount = 0;

readLock:
down(reader_can_try) //new sem blocks reader if writer waiting
down(rlock)
readCount++
if(readCount ==1)
  down(writer_can_enter) //don't coexist with a writer
up(rlock)
up(reader_can_try)

readUnlock:
down(rlock)
readCount--
if(readCount == 0)
  up(writer_can_enter)
up(rlock)

writeLock:
down(wlock)
writerCount++
if(writerCount==1)
  down(reader_can_try)
up(wlock)
down(writer_can_enter)  //release wlock and then block

writeUnlock:
down(wlock)
writerCount--
if(writerCount == 0)
  up(reader_can_try)
up(wlock)

up(writer_can_enter)
```

32. Consider the famous dining philosophers' problem. N philosophers are sitting around a table with N forks between them. Each philosopher must pick up both forks on her left and right before she can start eating. If each philosopher first picks the fork on her left (or right), then all will deadlock while waiting for the other fork. The goal is to come up with an algorithm that lets all philosophers eat, without deadlock or starvation. Write a solution to this problem using condition variables.

    Ans: A variable `state` is associated with each philosopher, and can be one of EATING (holding both forks) or THINKING (when not eating). Further, a condition variable is associated with each philosopher to make them sleep and wake them up when needed. Each philosopher must call the `pickup` function before eating, and `putdown` function when done. Both these functions use a mutex to change states only when both forks are available.

```
bothForksFree(i):
return (state[leftNbr(i)] != EATING &&
        state[rightNbr(i)] != EATING)

pickup(i):
  lock(mutex)
  while(!bothForksFree(i))
      wait(condvar[i])
  state[i] = EATING
  unlock(mutex)

putdown(i):
  lock(mutex)
  state[i] = THINKING
  if(bothForksFree(leftNbr(i)))
      signal(leftNbr(i))
  if(bothForksFree(rightNbr(i)))
      signal(rightNbr(i))
  unlock(mutex)
```

33. Consider a clinic with one doctor and a very large waiting room (of infinite capacity). Any patient entering the clinic will wait in the waiting room until the doctor is free to see her. Similarly, the the doctor also waits for a patient to arrive to treat. All communication between the patients and the doctor happens via a shared memory buffer. Any of the several patient processes, or the doctor process can write to it. Once the patient "enters the doctors office", she conveys her symptoms to the doctor using a call to `consultDoctor()`, which updates the shared memory with the patient's symptoms. The doctor then calls `treatPatient()` to access the buffer and update it with details of the treatment. Finally, the patient process must call `noteTreatment()` to see the updated treatment details in the shared buffer, before leaving the doctor's office. A template code for the patient and doctor processes is shown below. Enhance this code to correctly synchronize between the patient and the doctor processes. Your code should ensure that no race conditions occur due to several patients overwriting the shared buffer concurrently. Similarly, you must ensure that the doctor accesses the buffer only when there is valid new patient information in it, and the patient sees the treatment only after the doctor has written it to the buffer. You must use **only semaphores** to solve this problem. Clearly list the semaphore variables you use and their initial values first. Please pick sensible names for your variables.

    Ans:

    (a) Semaphores variables:

    ```
    pt_waiting = 0
    treatment_done = 0
    doc_avlbl = 1
    ```

    (b) Patient process:

    ```
    down(doc_avlbl)
    consultDoctor()
    up(pt_waiting)
    down(treatment_done)
    noteTreatment()
    up(doc_avlbl)
    ```

    (c) Doctor:

    ```
    while(1) {
    down(pt_waiting)
    treatPatient()
    up(treatment_done)
    }
    ```

34. Consider a multithreaded banking application. The main process receives requests to tranfer money from one account to the other, and each request is handled by a separate worker thread in the application. All threads access shared data of all user bank accounts. Bank accounts are represented by a unique integer account number, a balance, and a lock of type `mylock` (much like a `pthreads` mutex) as shown below.

```
struct account {
int accountnum;
int balance;
mylock lock;
};
```

Each thread that receives a transfer request must implement the transfer function shown below, which transfers money from one account to the other. Add correct locking (by calling the `dolock(&lock)` and `unlock(&lock)` functions on a `mylock` variable) to the tranfer function below, so that no race conditions occur when several worker threads concurrently perform transfers. Note that you must use the fine-grained per account lock provided as part of the account object itself, and not a global lock of your own. Also make sure your solution is deadlock free, when multiple threads access the same pair of accounts concurrently.

```
void transfer(struct account *from, struct account *to, int amount) {

from->balance -= amount; // dont write anything...
to->balance += amount; // ...between these two lines

}
```

Ans: The accounts must be locked in order of their account numbers. Otherwise, a transfer from account X to Y and a parallel transfer from Y to X may acquire locks on X and Y in different orders and end up in a deadlock.

```
struct account *lower = (from->accountnum < to->accountnum)?from:to;
struct account *higher = (from->accountnum < to->accountnum)?to:from;
dolock(&(lower->lock));
dolock(&(higher->lock));

from->balance -= amount;
to->balance += amount;

unlock(&(lower->lock));
unlock(&(higher->lock));
```

35. Consider a process with three threads A, B, and C. The default thread of the process receives multiple requests, and places them in a request queue that is accessible by all the three threads A, B, and C. For each request, we require that the request must first be processed by thread A, then B, then C, then B again, and finally by A before it can be removed and discarded from the queue. Thread A must read the next request from the queue only after it is finished with all the above steps of the previous one. Write down code for the functions run by the threads A, B, and C, to enable this synchronization. You can only worry about the synchronization logic and ignore the application specific processing done by the threads. You may use any synchronization primitive of your choice to solve this question.

    Ans: Solution using semaphores shown below. The order of processing is A1–B1–C–B2–A2. All threads run in a forever loop, and wait as dictated by the semaphores.

```
sem a1done = 0; b1done = 0; cdone = 0; b2done = 0;

ThreadA:
  get request from queue and process
  up(a1done)
  down(b2 done)
  finish with request

ThreadB:
  down(a1done)
  //do work
  up(b1done)
  down(cdone)
  //do work
  up(b2done)

ThreadC:
  down(b1done)
  //do work
  up(cdone)
```

36. Consider two threads A and B that perform two operations each. Let the operations of thread A be A1 and A2; let the operations of thread B be B1 and B2. We require that threads A and B each perform their first operation before either can proceed to the second operation. That is, we require that A1 be run before B2 and B1 before A2. Consider the following solutions based on semaphores for this problem (the code run by threads A and B is shown in two columns next to each other). For each solution, explain whether the solution is correct or not. If it is incorrect, you must also point out why the solution is incorrect.

(a)
```
sem A1Done = 0;  sem B1Done = 0;
//Thread A              //Thread B
A1                      B1
down(B1Done)            down(A1Done)
up(A1Done)              up(B1Done)
A2                      B2
```

(b)
```
sem A1Done = 0;  sem B1Done = 0;
//Thread A              //Thread B
A1                      B1
down(B1Done)            up(B1Done)
up(A1Done)              down(A1Done)
A2                      B2
```

(c)
```
sem A1Done = 0;  sem B1Done = 0;
//Thread A              //Thread B
A1                      B1
up(A1Done)              up(B1Done)
down(B1Done)            down(A1Done)
A2                      B2
```

Ans:

(a) Deadlocks, so incorrect.

(b) Correct

(c) Correct

37. Now consider a generalization of the above problem for the case of N threads that want to each execute their first operation before any thread proceeds to the second operation. Below is the code that each thread runs in order to achieve this synchronization. `count` is an integer shared variable, and `mutex` is a mutex binary semaphore that protects this shared variable. `step1Done` is a semaphore initialized to zero. You are told that this code is wrong and does not work correctly. Further, you can fix it by changing it slightly (e.g., adding one statement, or rearranging the code in some way). Suggest the change to be made to the code in the snippet below to fix it. You must use only semaphores and no other synchronization mechanism.

```
//run first step

down(mutex);
count++;
up(mutex);
if(count == N)
       up(step1Done);
down(step1Done);

//run second step
```

Ans: The problem is that the semaphore is decremented N times, but is only incremented once. To fix it, we must do up N times when count is N. Or, add up after the last down, so that it is performed N times by the N threads.

38. The cigarette smokers problem is a classical synchronization problem that involves 4 threads: one agent and three smokers. The smokers require three ingredients to smoke a cigarette: tobacco, paper, and matches. Each smoker has one of the three ingredients and waits for the other two, smokes the cigar once he obtains all ingredients, and repeats this forever. The agent repeatedly puts out two ingredients at a time and makes them available. In the correct solution of this problem, the smoker with the complementary ingredient should finish smoking his cigar. Consider the following solution to the problem. The shared variables are three semaphores `tobacco`, `paper` and `matches` initialized to 0, and semaphore `doneSmoking` initialized to 1. The agent code performs `down(doneSmoking)`, then picks two of the three ingredients at random and performs `up` on the corresponding two semaphores, and repeats. The smoker with tobacco runs the following code in a loop.

```
down(paper)
down(matches)
//make and smoke cigar
up(doneSmoking)
```

Similarly, the smoker with matches waits for tobacco and paper, and the smoker with paper waits for tobacco and matches, before signaling the agent that they are done smoking. Does the code above solve the synchronization problem correctly? If you answer yes, provide a justification for why the code is correct. If you answer no, describe what the error is and also provide a correct solution to the problem. (If you think the code is incorrect and are providing another solution, you may change the code of both the agent and the smokers. You can also introduce new variables as necessary. You must use only semaphores to solve the problem.)

Ans: The code is incorrect and deadlocks. One fix is to add semaphores for two ingredients at a time (e.g., tobaccoAndPaper). The smokers wait on these and the agent signals these. So there is no possibility of deadlock.

39. Consider a server program running in an online market place firm. The program receives buy and sell orders for one type of commodity from external clients. For every buy or sell request received by the server, the main process spawns a new buy or sell thread. We require that every buy thread waits until a sell thread arrives, and vice versa. A matched pair of buy and sell threads will both return a response to the clients and exit. You may assume that all buy/sell requests are identical to each other, so that any buy thread can be matched with any sell thread. The code executed by the buy thread is shown below (the code of the sell thread would be symmetric). You have to write the synchronization logic that must be run at the start of the execution of the thread to enable it to wait for a matching sell thread to arrive (if none exists already). Once the threads are matched, you may assume that the function `completeBuy()` takes care of the application logic for exchanging information with the matching thread, communicating with the client, and finishing the transaction. You may use any synchronization technique of your choice.

```
//declare any variables here

buy_thread_function:
  //start of sync logic


  //end of sync logic
  completeBuy();
```

Ans:

```
sem buyer = 0; sem seller = 0;

Buyer thread:

up(buyer)
down(seller)
completeBuy()
```

40. Consider the following classical synchronization problem called the barbershop problem. A barbershop consists of a room with N chairs. If a customer enters the barbershop and all chairs are occupied, then the customer leaves the shop. If the barber is busy, but chairs are available, then the customer sits in one of the free chairs and awaits his turn. The barber moves onto the next waiting seated customer after he finishes one hair cut. If there are no customers to be served, the barber goes to sleep. If the barber is asleep when a customer arrives, the customer wakes up the barber to give him a hair cut. A waiting customer vacates his chair after his hair cut completes. Your goal is to write the pseudocode for the customer and barber threads below with suitable synchronization. You must use only semaphores to solve this problem. Use the standard notation of invoking up/down functions on a semaphore variable.

    The following variables (3 semaphores and a count) are provided to you for your solution. You must use these variables and declare any additional variables if required.

    ```
    semaphore mutex = 1, customers = 0, barber = 0;
    int waiting_count = 0;
    ```

    Some functions to invoke in your customer and barber threads are:

    - A customer who finds the waiting room full should call the function `leave()` to exit the shop permanently. This function does not return.
    - A customer should invoke the function `getHairCut()` in order to get his hair cut. This function returns when the hair cut completes.
    - The barber thread should call `cutHair()` to give a hair cut. When the barber invokes this function, there should be exactly one customer invoking `getHairCut()` concurrently.

Ans:

```
Customer:

down(mutex)
if(waiting_count == N)
  up(mutex)
  leave()
waiting_count++
up(mutex)

up(customers)
down(barber)

getHairCut()

down(mutex)
waiting_count--
up(mutex)

Barber:

up(barber)
down(customers)
cutHair()
```

41. Consider a multithreaded application server handling requests from clients. Every new request that arrives at the server causes a new thread to be spawned to handle that request. The server can provide service to only one request/thread at a time, and other threads that arrive when the server is busy must wait for service using a synchronization primitive (semaphore or condition variable). In order to avoid excessive waiting times, the server does not wish to have more than N requests/threads in the system (including the waiting requests and any request it is currently serving). You may assume that $N > 2$. Given this constraint, a newly arriving thread must first check if N other requests are already in the system: if yes, it must exit without waiting and return an error value to the client, by calling the function `thr_exit_failure()`. This function terminates the thread and does not return.

    When a thread is ready for service, it must call the function `get_service()`. Your code should ensure that no more than one thread calls this function at any point of time. This function blocks the thread for the duration of the service. Note that, while the thread receiving service is blocked, other arriving threads must be free to join the queue, or exit if the system is overloaded. After a thread returns from `get_service()`, it must enable one of the waiting threads to seek service (if any are waiting), and then terminate itself succesfully by calling the function `thr_exit_success()`. This function terminates the thread and does not return.

    You are required to write pseudocode of the function to be run by the request threads in this system, as per the specification above. Your solution must use only locks and condition variables for synchronization. Clearly state all the variables used and their initial values at the start of your solution.

Ans

```
int num_requests=0;
bool server_busy = false
cv, mutex

lock(mutex)

if(num_requests == N)
  unlock(mutex)
  the_exit_failure()

num_requests++

if(server_busy)
  wait(cv, mutex)

server_busy = true
unlock(mutex)

get_service()

lock(mutex)
num_requests--
server_busy = false

if(num_requests > 0)
  signal(cv)

unlock(mutex)
thr_exit_success()
```

42. Consider the previous problem, but now assume that N is infinity. That is, all arriving threads will wait (if needed) for their turn in the queue of a synchronization primitive, get served when their turn comes, and exit successfully. Write the pseudocode of the function to be run by the threads with this modified specification. Your solution must only use semaphores for synchronization, and only the correct solution that uses the least number of semaphores will get full credit. Clearly state all the variables used and their initial values at the start of your solution.

Ans

```
sem waiting = 1

down(waiting)
get_service()
up(waiting)
thr_exit_success()
```

43. Consider the following synchronization problem. A group of children are picking chocolates from a box that can hold up to N chocolates. A child that wants to eat a chocolate picks one from the box to eat, unless the box is empty. If a child finds the box to be empty, she wakes up the mother, and waits until the mother refills the box with N chocolates. Unsynchronized code snippets for the child and mother threads are as shown below:

```
//Child
while True:
  getChocolateFromBox()
  eat()

//Mother
while True:
  refillChocolateBox(N)
```

You must now modify the code of the mother and child threads by adding suitable synchronization such that a child invokes `getChocolateFromBox()` only if the box is non-empty, and the mother invokes `refillChocolateBox(N)` only if the box is fully empty. Solve this question using only locks and condition variables, and no other synchronization primitive. The following variables have been declared for use in your solution.

```
int count = 0;
mutex m; // you may invoke lock and unlock
condvar fullBox, emptyBox; //you may perform wait and signal
                           //or signal_broadcast
```

(a) Code for child thread

(b) Code for mother thread

Ans:

```
//Child
while True:
  lock(m)
  while(count == 0)
    signal(emptyBox)
    wait(fullBox, m)
  getChocolateFromBox()
  eat()
  count--
  signal(fullBox) //optional
  unlock(m)

//Mother
while True:
  lock(m)
  if(count > 0)
    wait(emptyBox, m)
  refillChocolateBox(N)
  count += N
  signal(fullBox)
  unlock(m)
```

There are two ways of waking up sleeping children. Either the mother does a signal broadcast to all children. Or every child that eats a chocolate wakes up another sleeping child. You may also assume that signal by mother wakes up all children.

44. Repeat the above question, but your solution now must use only semaphores and no other synchronization primitive. The following variables have been declared for use in your solution.

```
int count = 0;
semaphore m, fullBox, emptyBox;
//initial values of semaphores are not specified
//you may invoke up and down methods on a semaphore
```

(a) Initial values of the semaphores

(b) Code for child thread

(c) Code for mother thread

Ans:

```
m = 1, fullBox = 0, emptyBox = 0

//Child
while True:
  down(m)
  if(count == 0)
    up(emptyBox)
    down(fullBox)
    count += N
  getChocolateFromBox()
  eat()
  count--
  up(m)

//Mother
while True:
  down(emptyBox)
  refillChocolateBox(N)
  up(fullBox)
```

Here the subtlety is the lock m. Mother can't get lock to update count after filling the box, as that will cause a deadlock. In general, if child sleeps with mutex m locked, then mother cannot request the same lock.

45. Consider the classic "barrier" synchronization problem, where N threads wish to synchronize with each other as follows. N threads arrive into the system at different times and in any order. The arriving threads must wait until all N threads have arrived into the system, and continue execution only after all N threads have arrived. We wish to write logic to synchronize the threads in the manner stated above using semaphores. Below are three possible solutions to the problem. You are told that one of the solutions is correct and the other two are wrong. Identify the correct solution amongst the three given options. Further, for each of the other incorrect solutions, explain clearly why the solution is wrong. The following shared variables are declared for use in each solution.

```
int count = 0;
sem mutex; //initialized to 1
sem barrier; //initialized to 0
```

(a)
```
down(mutex)
  count++
  if(count == N) up(barrier)
up(mutex)

down(barrier)

//wait done; proceed to actual task
```

(b)
```
down(mutex)
  count++
  if(count == N) up(barrier)
up(mutex)

down(barrier)
up(barrier)

//wait done; proceed to actual task
```

(c)
```
down(mutex)
  count++
  if(count == N) up(barrier)
  down(barrier)
  up(barrier)
up(mutex)

//wait done; proceed to actual task
```

Ans: In (a) up is done only once when many threads are waiting on down. In (c), down(barrier) is called when mutex held, so code deadlocks. (b) is correct answer.

46. Consider the barrier synchronization primitive discussed in class, where the N threads of an application wait until all the threads have arrived at a barrier, before they proceed to do a certain task. You are now required to write the code for a reusable barrier, where the N application threads perform a series of steps in a loop, and use the same barrier code to synchronize for each iteration of the loop. That is, your solution should ensure that all threads wait for each other before the start of each step, and proceed to the next step only after all threads have completed the previous step. Your solution must only use semaphores. The following functions can be invoked on a semaphore `s` used in this question: `down(s)`, `up(s)`, and `up(s,n)`. While the first two functions are as studied in class, the function `up(s,n)` simply invokes `up(s)` n times atomically.

    We have provided you some code to get started. Shown below is the code to be run by each application thread, including the code to wait at the barrier. However, this is not the correct solution, as this code only works as a single-use barrier, i.e., it only ensures that the threads synchronize at the barrier once, and cannot be used to synchronize multiple times (can you figure out why?). You are required to modify this code to make it reusable, such that the threads can synchronize at the barrier multiple times for the multiple steps to be performed.

    Your solution must only use the following variables: `int count = 0`; and semaphores (initial values as given): `sem mutex = 1`; `sem barrier1 = 0`; `sem barrier2 = 0`;

```
For each step to be executed by the threads, do:

//add code here if required to make barrier reusable


down(mutex)
  count++
  if(count == N) up(barrier1, N)
up(mutex)
down(barrier1)

... wait done, execute actual task of this step ...

//add code here if required to make barrier reusable for next step
```

    Ans: The extra code to be added is at the end of completing a step, where you make all threads wait once again.

```
down(mutex)
count--
if(count==0) up(barrier2, N)
up(mutex)
down(barrier2)
```

47. Consider a web server that is supposed to serve a batch of N requests. Each request that arrives at the web server spawns a new thread. The arriving threads wait until N of them accumulate, at which point all of them proceed to get service from the server. Shown below is the code executed by each arriving thread, that causes it to wait until all the other threads arrive. The variable `count` is initialized to N. The code also uses `wait` and `signal` primitives on a condition variable; and you may assume that the signal primitive wakes up all waiting threads (not just one of them).

```
lock(mutex)
  count--;
unlock(mutex)

if(count > 0) {
  lock(mutex)
  wait(cv, mutex)
  unlock(mutex)
}
else {
  lock(mutex)
  signal(cv)
  unlock(mutex)
}

... wait done, proceed to server ...
```

You are told that the code above is incorrect, and can sometimes cause a deadlock. That is, in some executions, all N threads do not go to the server for service, even though they have arrived.

(a) Using an example, explain the exact sequence of events that can cause a deadlock. You must write your answers as bullet points, with one event per bullet point, starting from threads arriving in the system until the deadlock.

(b) Explain how you will fix this deadlock and correct the code shown above. You must retain the basic structure of the code. Indicate your changes next to the code snippet above.

Ans: The given incorrect solution may cause a missed wakeup. For example, some thread decides to wait and goes inside the if-loop, but is context switched out before calling wait (and before it acquires the lock). Now, if count hits 0 and signal happens before it runs again, it will wait with no one to wake it up, leading to deadlock. The fix is simply holding the lock all through the condition checking and waiting.

48. Consider an application that has $K + 1$ threads running on a Linux-like OS ($K > 1$). The first K threads of an application execute a certain task T1, and the remaining one thread executes task T2. The application logic requires that task T1 is executed $N > 1$ times, followed by task T2 executed once, and this cycle of N executions of T1 followed by one execution of T2 continue indefinitely. All K threads should be able to participate in the N executions of task T1, even though it is not required to ensure perfect fairness amongst the threads.

    Shown below is one possible set of functions executed by the threads running tasks T1 and T2. You are told that this solution has two bugs in the code run by the thread performing task T2. Briefly describe the bugs in the space below, and suggest small changes to the corresponding code to fix these bugs (you may write your changes next to the code snippet). You must not change the code corresponding to task T1 in any way. All threads share a counter `count` (initialized to 0), a mutex variable `m`, and two condition variables `t1cv`, and `t2cv`. Here, the function `signal` on a condition variable wakes up only one of the possibly many sleeping threads.

```
//function run by K threads of task T1
while True {
    lock(m)
    if(count >= N) {
        signal(t2cv)
        wait(t1cv, m)
    }
    //.. do task T1 once ..
    count++
    unlock(m)
}
//function run by thread of task T2
while True {
    lock(m)
    wait(t2cv, m)
    // .. do task T2 once
    count = 0
    signal(t1cv)
    unlock(m)
}
```

    Ans: (a) check count<N and only then wait (b) signal broadcast instead of signal

49. You are now required to solve the previous question using semaphores for synchronization. You are given the pseudocode for the function run by the thread executing task T2 (which you must not change). You are now required to write the corresponding code executed by the K threads running task T1. You must use the following semaphores in your solution: `mutex`, `t1sem`, `t2sem`. You must initialize them suitably below. The variable `count` (initialized to 0) is also available for use in your solution.

Ans:

```
//fill in initial values of semaphores
sem_init(mutex,   );  sem_init(t1sem,   );   sem_init(t2sem,    );
//other variables
int count = 0

//function run by thread executing T2
while True {
    down(t2sem)
    //.. do task T2 ..
    up(t1sem)
}

//function run by threads executing task T1
while True {


}
```

Ans:

```
mutex=1, t1sem=0, t2sem=0

down(mutex)
if(count == N)
  up(t2sem)
  down(t1sem)
  count = 0

do task T1 once
count++
up(mutex)
```

50. Multiple people are entering and exiting a room that has a light switch. You are writing a computer program to model the people in this situation as threads in an application. You must fill in the functions `onEnter()` and `onExit()` that are invoked by a thread/person when the person enters and exits a room respectively. We require that the first person entering a room must turn on the light switch by invoking the function `turnOnSwitch()`, while the last person leaving the room must turn off the switch by invoking `turnOffSwitch()`. You must invoke these functions suitably in your code below. You may use any synchronization primitives of your choice to achieve this desired goal. You may also use any variables required in your solution, which are shared across all threads/persons.

    (a) Variables and initial values

    (b) Code `onEnter()` to be run by thread/person entering

    (c) Code `onExit()` to be run by thread/person exiting

    Ans:

```
variables: mutex, count

onEnter():
lock(mutex)
count++
if(count==1) turnOnSwitch()
unlock(mutex)

onExit():
lock(mutex)
count--
if(count==0) turnOffSwitch()
unlock(mutex)
```

Lectures on Operating Systems (Mythili Vutukuru, IIT Bombay)

Practice Problems: File systems

1. Provide one reason why a DMA-enabled device driver usually gives better performance over a non-DMA interrupt-driven device driver.

   Ans: A DMA driver frees up CPU cycles that would have been spent copying data from the device to physical memory.

2. Which of the following statements is/are true regarding memory-mapped I/O?

   A. The CPU accesses the device memory much like it accesses main memory.

   B. The CPU uses separate architecture-specific instructions to access memory in the device.

   C. Memory-mapped I/O cannot be used with a polling-based device driver.

   D. Memory-mapped I/O can be used only with an interrupt-driven device driver.

   Ans: A

3. Consider a file D1/F1 that is hard linked from another parent directory D2. Then the directory entry of this file (including the filename and inode number) in directory D1 must be exactly identical to the directory entry in directory D2. [T/F]

   Ans: F (the file name can be different)

4. It is possible for a system that uses a disk buffer cache with FIFO as the buffer replacement policy to suffer from the Belady's anomaly. [T/F]

   Ans: T

5. Reading files via memory mapping them avoids an extra copy of file data from kernel space buffers to user space buffers. [T/F]

   Ans: T

6. A soft link can create a link between files across different file systems, whereas a hard link can only create links between a directory and a file within the same file system. [T/F]

   Ans: T (becasue hard link stores inode number, which is unique only within a file system)

7. Consider the process of opening a new file that does not exist (obviously, creating it during opening), via the "open" system call. Describe changes to all the in-memory and disk-based file system structures (e.g., file tables, inodes, and directories) that occur as part of this system call implementation. Write clearly, listing the structure that is changed, and the change made to it.

   Ans: (a) New inode allocated on disk (with link count=1), and inode bitmap updated in the process. (b) Directory entry added to parent directory, to add mapping from file name to inode number.

(c) In-memory inode allocated. (d) System-wide open file table points to in-memory inode. (e) Per-process file descriptor table points to open file table entry.

8. Now, suppose the process that has opened the file in the previous question proceeds to write 100 bytes into the file. Assume block size on disk is 512 bytes. Assume the OS uses a write-through disk buffer cache. List all the operations/changes to various datastructures that take place when the write operation successfully completes.

   Ans: (a) open file table offset is changed (b) in-memory and on-disk inode adds pointer to new data block, and last modified time is updated (c) a copy of the data block comes into the disk buffer cache (d) New data block is allocated from data block bitmap (e) new data block is filled with user provided data

9. Repeat the above question for the implementation of the "link" system call, when linking to an existing file (not open from any process) in a directory from another new parent directory.

   Ans: (a) The link count of the on-disk inode of the file is incremented. (b) A directory entry is added to the new directory to create a mapping from the file name to the inode number of the original file (if the new directory does not have space in its data blocks for the new file, a new data block is allocated for the new directory entry, and a pointer to this data block is added from the directory's inode).

10. Repeat the above question for the implementation of the "dup" system call on a file descriptor.

    Ans: To dup a file descriptor, another empty slot in the file descriptor table of the process is found, and this new entry is set to point to the same global open file table entry as the old file descriptor. That is, two FDs point to same system-wide file table entry.

11. Consider a file system with 512-byte blocks. Assume an inode of a file holds pointers to N direct data blocks, and a pointer to a single indirect block. Further, assume that the single indirect block can hold pointers to M other data blocks. What is the maximum file size that can be supported by such an inode design?

    Ans: (N+M)*512 bytes

12. Consider a FAT file system where disk is divided into M byte blocks, and every FAT entry can store an N bit block number. What is the maximum size of a disk partition that can be managed by such a FAT design?

    Ans: $2^N * M$ bytes

13. Consider a secondary storage system of size 2 TB, with 512-byte sized blocks. Assume that the filesystem uses a multilevel inode datastructure to track data blocks of a file. The inode has 64 bytes of space available to store pointers to data blocks, including a single indirect block, a double indirect block, and several direct blocks. What is the maximum file size that can be stored in such a file system?

    Ans: Number of data blocks = $2^{41}/2^9 = 2^{32}$, so 32 bits or 4 bytes are required to store the number of a data block.

    Number of data block pointers in the inode = 64/4 = 16, of which 14 are direct blocks. The single indirect block stores pointers to 512/4 = 128 data blocks. The double indirect block points to 128 single indirect blocks, which in turn point to 128 data blocks each.

So, the total number of data blocks in a file can be 14 + 128 + 128*128 = 16526, and the maximum file size is 16526*512 bytes.

14. Consider a filesystem managing a disk with block size 2^b bytes, and disk block addresses of 2^a bytes. The inode of a file contains n direct blocks, one single indirect block, one double indirect block, and one triple indirect block. What is the maximum size of a file (in bytes) that can be stored in this filesystem? Assume that the indirect blocks only store a sequence of disk addresses, and no other metadata.

    Ans: Let x = number of disk addresses per block = 2^{b-a}. Then max file size is $2^b * (n + x + x^2 + x^3)$.

15. The `fork` system call creates new entries in the open file table for the newly created child process. [T/F]

    Ans: F

16. When a process opens a file that is already being read by another process, the file descriptors in both processes will point to the same open file table entry. [T/F]

    Ans: F

17. Memory mapping a file using the `mmap` system call adds one or more entries to the page table of the process. [T/F]

    Ans: T

18. The read system call to fetch data from a file always blocks the invoking process. [T/F]

    Ans: F (the data may be readily available in the disk buffer cache)

19. During filesystem operations, if the filesystem implementation ensures that changes to data blocks of a file are flushed to disk before changes to metadata blocks (like inodes and bitmaps), then the filesystem will never be in an inconsistent state after a crash, and a filesystem checker need not be run to detect and fix any inconsistencies. [T/F]

    Ans: F (If there are multiple metadata operations, some may have happened and some may have been lost, causing an inconsistency. For example, a bitmap may indicate a data block is allocated but no inode points to it.)

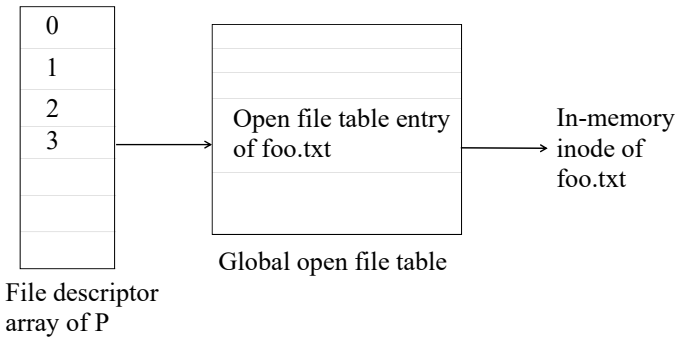
20. Interrupt-based device drivers give superior performance to polling-based drivers because they eliminate the time spent by the CPU in copying data to and from the device hardware. [T/F]

    Ans: F

21. When a process writes a block to the disk via a disk buffer cache using the write-back policy, the process invoking the write will block until the write is committed to disk. [T/F]

    Ans: F

22. Consider a process P that has opened a file `foo.txt` using the `open` system call. The figure below shows the file descriptor array of P and the global open file table, and the pointers linking these data structures.



(a) After opening the file, P forks a child C. Draw a figure showing the file descriptor arrays of P and C, and the global open file table, immediately after the fork system call successfully completes. It is enough to show the entries pertaining to the file `foo.txt`, as in the figure above.

(b) Repeat part (a) for the following scenario: after P forks a child C, another process Q also opens the same file `foo.txt`.

Ans: In (a), the file descriptor arrays of P and C are pointing to the same file table entry. In (b), the file descriptor array of Q is pointing to a new open file table entry, which points to the same inode of the file foo.txt.

Lectures on Operating Systems (Mythili Vutukuru, IIT Bombay)

Practice Problems: Memory

1. Provide one advantage of using the slab allocator in Linux to allocate kernel objects, instead of simply allocating them from a dynamic memory heap.

   Ans: A slab allocator is fast because memory is preallocated. Further, it avoids fragmentation of kernel memory.

2. In a 32-bit architecture machine running Linux, for every physical memory address in RAM, there are at least 2 virtual addresses pointing to it. That is, every physical address is mapped at least twice into the virtual address space of some set of processes. [T/F]

   Ans: F (this may be true of simple OS like xv6 studied in class, but not generally true)

3. Consider a system with N bytes of physical RAM, and M bytes of virtual address space per process. Pages and frames are K bytes in size. Every page table entry is P bytes in size, accounting for the extra flags required and such. Calculate the size of the page table of a process.

   Ans: M/K * P

4. The memory addresses generated by the CPU when executing instructions of a process are called logical addresses. [T/F]

   Ans: T

5. When a C++ executable is run on a Linux machine, the kernel code is part of the executable generated during the compilation process. [T/F]

   Ans: F (it is only part of the virtual address space)

6. When a C++ executable is run on a Linux machine, the kernel code is part of the virtual address space of the running process. [T/F]

   Ans: T

7. Consider a Linux-like OS running on x86 Intel CPUs. Which of the following events requires the OS to update the page table pointer in the MMU (and flush the changes to the TLB)? Answer "update" or "no update".

   (a) A process moves from user mode to kernel mode.

   Ans: no update

   (b) The OS switches context from one process to another.

   Ans: update

8. Consider a process that has just forked a child. The OS implements a copy-on-write fork. At the end of the fork system call, the OS does not perform a context switch and will return back to the user mode of the parent process. Now, which of the following entities are updated at the end of a successful implementation of the fork system call? Answer "update" or "no update".

   (a) The page table of the parent process.
   Ans: update because the parent's pages must be marked read-only.

   (b) The page table information in the MMU and the TLB.
   Ans: update because the parent's pages must be marked read-only.

9. A certain page table entry in the page table of a process has both the valid and present bits set. Describe what happens on a memory access to a virtual address belonging to this page table entry.

   (a) What happens at the TLB? (hit/miss/cannot say)
   Ans: cannot say

   (b) Will a page fault occur? (yes/no/cannot say)
   Ans no

10. A certain page table entry in the page table of a process has the valid bit set but the present bit unset. Describe what happens on a memory access to a virtual address belonging to this page table entry.

    (a) What happens at the TLB? (hit/miss/cannot say)
    Ans: miss

    (b) Will a page fault occur? (yes/no/cannot say)
    Ans yes

11. Consider the page table entries within the page table of a process that map to kernel code/data stored in RAM, in a Linux-like OS studied in class.

    (a) Are the physical addresses of the kernel code/data stored in the page tables of various process always the same? (yes/no/cannot say)
    Ans: yes, because there is only one copy of kernel code in RAM

    (b) Does the page table of every process have page table entries pointing to the kernel code/data? (yes/no/cannot say)
    Ans: yes, because every process needs to run kernel code in kernel mode

12. Consider a process P running in a Linux-like operating system that implements demand paging. The page/frame size in the system is 4KB. The process has 4 pages in its heap. The process stores an array of 4K integers (size of integer is 4 bytes) in these 4 pages. The process then proceeds to access the integers in the array sequentially. Assume that none of these 4 pages of the heap are initially in physical memory. The memory allocation policy of the OS allocates only 3 physical frames at any point of time, to the store these 4 pages of the heap. In case of a page fault and all 3 frames have been allocated to the heap of the process, the OS uses a LRU policy to evict one of these 4 pages to make space for the new page. Approximately what fraction of the 4K accesses to array elements will result in a page fault?

(a) Almost 100

(b) Approximately 25

(c) Approximately 75

(d) Approximately 0.1

Ans: (d)

13. Given below are descriptions of different entries in the page table of a process, with respect to which bits are set and which are not set. Accessing which of the page table entries below will always result in the MMU generating a trap to the OS during address translation?

    (a) Page with both valid and present bits set

    (b) Page with valid bit set but present bit unset

    (c) Page with valid bit unset

    (d) Page with valid, present, and dirty bits set

    Ans: (b), (c)

14. Which of the following statements is/are true regarding the memory image of a process?

    (a) Memory for non-static local variables of a function is allocated on the heap dynamically at run time

    (b) Memory for arguments to a function is allocated on the stack dynamically at run time

    (c) Memory for static and global variables is allocated on the stack dynamically at run time

    (d) Memory for the argc, argv arguments to the main function is allocated in the code/data section of the executable at compile time

    Ans: (b)

15. Consider a process P in a Linux-like operating system that implements demand paging. Which of the following pages in the page table of the process will have the valid bit set but the present bit unset?

    (a) Pages that have been used in the past by the process, but were evicted to swap space by the OS due to memory pressure

    (b) Pages that have been requested by the user using mmap/brk/sbrk system calls, but have not yet been accessed by the user, and hence not allocated physical memory frames by the OS

    (c) Pages corresponding to unused virtual addresses in the virtual address space of the process

    (d) Pages with high virtual addresses mapping to OS code and data

    Ans: (a), (b)

16. Consider a process P in a Linux-like operating system that implements demand paging. For a particular page in the page table of this process, the valid and present bits are both set. Which of the following are possible outcomes that can happen when the CPU accesses a virtual address in this page of the process? Select all outcomes that are possible.

(a) TLB hit (virtual address found in TLB)

(b) TLB miss (virtual address not found in TLB)

(c) MMU walks the page table (to translate the address)

(d) MMU traps to the OS (due to illegal access)

Ans: (a), (b), (c), (d)

17. Consider a process P in a Linux-like operating system that implements demand paging using the LRU page replacement policy. You are told that the i-th page in the page table of the process has the accessed bit set. Which of the following statements is/are true?

(a) This bit was set by OS when it allocated a physical memory frame to the page

(b) This bit was set by MMU when the page was accessed in the recent past

(c) This page is likely to be evicted by the OS page replacement policy in the near future

(d) This page will always stay in physical memory as long as the process is alive

Ans: (b)

18. Which of the following statements is/are true regarding the functions of the OS and MMU in a modern computer system?

(a) The OS sets the address of the page table in a CPU register accessible to the MMU every time a new process is created in the system

(b) The OS sets the address of the page table in a CPU register accessible to the MMU every time a new process is context switched in by the CPU scheduler

(c) MMU traps to OS every time an address is not found in the TLB cache

(d) MMU traps to OS every time it cannot translate an address using the page table available to it

Ans: (b), (d)

19. Consider a modern computer system using virtual addressing and translation via MMU. Which of the following statements is/are valid advantages of using virtual addressing as opposed to directly using physical addresses to fetch instructions and data from main memory?

(a) One does not need to know the actual addresses of instructions and data in main memory when generating compiled executables.

(b) One can easily provide isolation across processes by limiting the physical memory that is mapped into the virtual address space of a process.

(c) Using virtual addressing allows us to hide the fact that user's memory is allocated non-contiguously, and helps provide a simplified view to the user.

(d) Memory access using virtual addressing is faster than directly accessing memory using physical addresses.

Ans: (a), (b), (c)

20. Consider a process running on a system with a 52-bit CPU (i.e., virtual addresses are 48 bits in size). The system has a physical memory of 8GB. The page size in the system is 4KB, and the size of a page table entry is 4 bytes. The OS uses hierarchical paging. Which of the following statements is/are true? You can assume 2^{10} = 1K, 2^{20} = 1M, and so on.

    (a) We require a 4-level page table to keep track of the virtual address space of a process.

    (b) We require a 5-level page table to keep track of the virtual address space of a process.

    (c) The most significant 9 bits are used to index into the outermost page directory by the MMU during address translation.

    (d) The most significant 40 bits of a virtual address denote the page number, and the least significant 12 bits denote the offset within a page.

    Ans: (a), (d)

21. Consider the following line of code in a function of a process.
    ```
    int *x = (int *)malloc(10 * sizeof(int));
    ```
    When this function is invoked and executed:

    (a) Where is the memory for the variable `x` allocated within the memory image of the process? (stack/heap)

    Ans: stack

    (b) Where is the memory for the 10 integer variables allocated within the memory image of the process? (stack/heap)

    Ans: heap

22. Consider an OS that is not using a copy-on-write implementation for the `fork` system call. A process P has spawned a child C. Consider a virtual address v that is translated to physical address $A_p(v)$ using the page table of P, and to $A_c(v)$ using the page table of C.

    (a) For which virtual addresses v does the relationship $A_p(v) = A_c(v)$ hold?

    Ans: For kernel space addresses, shared libraries and such.

    (b) For which virtual addresses v does the relationship $A_p(v) = A_c(v)$ not hold?

    Ans: For userspace part of memory image, e.g., code, data, stack, heap.

23. Consider a system with paging-based memory management, whose architecture allows for a 4GB virtual address space for processes. The size of logical pages and physical frames is 4KB. The system has 8GB of physical RAM. The system allows a maximum of 1K (=1024) processes to run concurrently. Assuming the OS uses hierarchical paging, calculate the maximum memory space required to store the page tables of *all* processes in the system. Assume that each page table entry requires an additional 10 bits (beyond the frame number) to store various flags. Assume page table entries are rounded up to the nearest byte. Consider the memory required for both outer and inner page tables in your calculations.

    Ans:

    Number of physical frames = $2^{33}/2^{12} = 2^{21}$. Each PTE has frame number (21 bits) and flags (10 bits) $\approx$ 4 bytes. The total number of pages per process is $2^{32}/2^{12}$=2^{20}, so total size of inner page table pages is $2^{20} \times 4$ = 4MB.

Each page can hold $2^{12}/4 = 2^{10}$ PTEs, so we need $2^{20}/2^{10}$ PTEs to point to inner page tables, which will fit in a single outer page table. So the total size of page tables of one process is 4MB + 4KB. For 1K process, the total memory consumed by page tables is 4GB + 4MB.

24. Consider a simple system running a single process. The size of physical frames and logical pages is 16 bytes. The RAM can hold 3 physical frames. The virtual addresses of the process are 6 bits in size. The program generates the following 20 virtual address references as it runs on the CPU: `0, 1, 20, 2, 20, 21, 32, 31, 0, 60, 0, 0, 16, 1, 17, 18, 32, 31, 0, 61`. (Note: the 6-bit addresses are shown in decimal here.) Assume that the physical frames in RAM are initially empty and do not map to any logical page.

    (a) Translate the virtual addresses above to logical page numbers referenced by the process. That is, write down the reference string of 20 page numbers corresponding to the virtual address accesses above. Assume pages are numbered starting from 0, 1, ...

    (b) Calculate the number of page faults genrated by the accesses above, assuming a FIFO page replacement algorithm. You must also correctly point out which page accesses in the reference string shown by you in part (a) are responsible for the page faults.

    (c) Repeat (b) above for the LRU page replacement algorithm.

    (d) What would be the lowest number of page faults achievable in this example, assuming an optimal page replacement algorithm were to be used? Repeat (b) above for the optimal algorithm.

    Ans:

    (a) For 6 bit virtual addresses, and 4 bit page offsets (page size 16 bytes), the most significant 2 bits of a virtual address will represent the page number. So the reference string is 0, 0, 1, 0, 1, 1, 2, 1, 0, 3 (repeated again).

    (b) Page faults with FIFO = 8. Page faults on 0,1,2,3 (replaced 0), 0 (replaced 1), 1 (replaced 2), 2 (replaced 3), 3.

    (c) Page faults with LRU = 6. Page faults on 0, 1, 2, 3 (replaced 2), 2 (replaced 3), 3.

    (d) The optimum algorithm will replace the page least likely to be used in future, and would look like LRU above.

25. Consider a system with only virtual addresses, but no concept of virtual memory or demand paging. Define *total memory access time* as the time to access code/data from an address in physical memory, including the time to resolve the address (via the TLB or page tables) and the actual physical memory access itself. When a virtual address is resolved by the TLB, experiments on a machine have empirically observed the total memory access time to be (an approximately constant value of) t_h. Similarly, when the virtual address is not in the TLB, the total memory access time is observed to be t_m. If the average total memory access time of the system (averaged across all memory accesses, including TLB hits as well as misses) is observed to be t_x, calculate what fraction of memory addresses are resolved by the TLB. In other words, derive an expression for the TLB hit rate in terms of t_h, t_m, and t_x. You may assume $t_m > t_h$.

    Ans: We have $t_x = h * t_h + (1 - h) * t_m$, so $t_h = \frac{t_m - t_x}{t_m - t_h}$

26. 4. Consider a system with a 6 bit virtual address space, and 16 byte pages/frames. The mapping from virtual page numbers to physical frame numbers of a process is (0,8), (1,3), (2,11), and (3,1). Translate the following virtual addresses to physical addresses. Note that all addresses are in decimal. You may write your answer in decimal or binary.

    (a) 20

    (b) 40

    Ans:

    (a) 20 = 01 0100 = 11 0100 = 52

    (b) 40 = 10 1000 = 1011 1000 = 184

27. Consider a system with several running processes. The system is running a modern OS that uses virtual addresses and demand paging. It has been empirically observed that the memory access times in the system under various conditions are: t1 when the logical memory address is found in TLB cache, t2 when the address is not in TLB but does not cause a page fault, and t3 when the address results in a page fault. This memory access time includes all overheads like page fault servicing and logical-to-physical address translation. It has been observed that, on an average, 10% of the logical address accesses result in a page fault. Further, of the remaining virtual address accesses, two-thirds of them can be translated using the TLB cache, while one-third require walking the page tables. Using the information provided above, calculate the average expected memory access time in the system in terms of t1,t2, and t3.

    Ans: 0.6*t1 + 0.3*t2 + 0.1*t3

28. Consider a system where each process has a virtual address space of 2^v bytes. The physical address space of the system is 2^p bytes, and the page size is 2^k bytes. The size of each page table entry is 2^e bytes. The system uses hierarchical paging with l levels of page tables, where the page table entries in the last level point to the actual physical pages of the process. Assume $l \geq 2$. Let v_0 denote the number of (most significant) bits of the virtual address that are used as an index into the outermost page table during address translation.

    (a) What is the number of logical pages of a process?

    (b) What is the number of physical frames in the system?

    (c) What is the number of PTEs that can be stored in a page?

    (d) How many pages are required to store the innermost PTEs?

    (e) Derive an expression for l in terms of v, p, k, and e.

    (f) Derive an expression for v_0 in terms of l, v, p, k, and e.

    Ans:

    (a) 2^{v-k}

    (b) 2^{p-k}

    (c) 2^{k-e}

    (d) 2^{v-k} / 2^{k-e} = 2^{v+e-2k}

(e) The least significant k of v bits indicate offset within a page. Of the remaining $v - k$ bits, $k - e$ bits will be used to index into the page tables at every level, so the number of levels l = ceil $\frac{v-k}{k-e}$

(f) $v - k - (l - 1) * (k - e)$

29. Consider an operating system that uses 48-bit virtual addresses and 16KB pages. The system uses a hierarchical page table design to store all the page table entries of a process, and each page table entry is 4 bytes in size. What is the total number of pages that are required to store the page table entries of a process, across all levels of the hierarchical page table?

Ans: Page size = 2^{14} bytes. So, the number of page table entries = $2^{48}/2^{14} = 2^{34}$. Each page can store 16KB/4 = 2^{12} page table entries. So, the number of innermost pages = 2^{34} / $2^{12} = 2^{22}$.

Now, pointers to all these innermost pages must be stored in the next level of the page table, so the next level of the page table has 2^{22} / $2^{12} = 2^{10}$ pages. Finally, a single page can store all the 2^{10} page table entries, so the outermost level has one page.

So, the total number of pages that store page table entries is $2^{22} + 2^{10} + 1$.

30. Consider a memory allocator that uses the buddy allocation algorithm to satisfy memory requests. The allocator starts with a heap of size 4KB (4096 bytes). The following requests are made to the allocator by the user program (all sizes requested are in bytes): ptr1 = malloc(500); ptr2 = malloc(200); ptr3 = malloc(800); ptr4 = malloc(1500). Assume that the header added by the allocator is less than 10 bytes in size. You can make any assumption about the implementation of the buddy allocation algorithm that is consistent with the description in class.

(a) Draw a figure showing the status of the heap after these 4 allocations complete. Your figure must show which portions of the heap are assigned and which are free, including the sizes of the various allocated and free blocks.

(b) Now, suppose the user program frees up memory allocations of ptr2, ptr3, and ptr4. Draw a figure showing the status of the heap once again, after the memory is freed up and the allocation algorithm has had a chance to do any possible coalescing.

Ans:

(a) [512 B][256 B] 256 B free [1024 B][2048 B]

(b) [512 B] 512 B free, 1024 B free, 2048 B free. No further coalescing is possible.

31. Consider a system with 8-bit virtual and physical addresses, and 16 byte pages. A process in this system has 4 logical pages, which are mapped to 3 physical pages in the following manner: logical page 0 maps to physical page 6, 1 maps to 3, 2 maps to 11, and logical page 5 is not mapped to any physical page yet. All the other pages in the virtual address space of the process are marked invalid in the page table. The MMU is given a pointer to this page table for address translation. Further, the MMU has a small TLB cache that stores two entries, for logical pages 0 and 2. For each virtual address shown below, describe what happens when that address is accessed by the CPU. Specifically, you must answer what happens at the TLB (hit or miss?), MMU (which page table entry is accessed?), OS (is there a trap of any kind?), and the physical memory (which physical address is accessed?). You may write the translated physical address in binary format. (Note that it is not implied that the accesses below happen one after the other; you must solve each part of the question independently using the information provided above.)

(a) Virtual address 7

(b) Virtual address 20

(c) Virtual address 70

(d) Virtual address 80

Ans:

(a) 7 = 0000 (page number) + 0111 (offset) = logical page 0. TLB hit. No page table walk. No OS trap. Physical address 0110 0111 is accessed.

(b) 20 = 0001 0100 = logical page 1. TLB miss. MMU walks page table. Physical address 0011 0100

(c) 70 = 0100 0110 = logical page 4. TLB miss. MMU accesses page table and discovers it is an invalid entry. MMU raises trap to OS.

(d) 80 = 0101 0000 = logical page 5. TLB miss. MMU accesses page table and discovers page not present. MMU raises a page fault to the OS.

32. Consider a system with 8-bit addresses and 16-byte pages. A process in this system has 4 logical pages, which are mapped to 3 physical frames in the following manner: logical page 0 maps to physical frame 2, page 1 maps to frame 0, page 2 maps to frame 1, and page 3 is not mapped to any physical frame. The process may not use more than 3 physical frames. On a page fault, the demand paging system uses the LRU policy to evict a page. The MMU has a TLB cache that can store 2 entries. The TLB cache also uses the LRU policy to store the most recently used mappings in cache. Now, the process accesses the following logical addresses in order: 7, 17, 37, 20, 40, 60.

(a) Out of the 6 memory accesses, how many result in a TLB miss? Clearly indicate the accesses that result in a miss. Assume that the TLB cache is empty before the accesses begin.

Ans: 0,1,2, (miss) 1,2 (hit), 3 (miss)

(b) Out of the 6 memory accesses, how many result in a page fault? Clearly indicate the accesses that result in a page fault.

Ans: last access 3 result in a page fault

(c) Upon accessing the logical address 60, which physical address is eventually accessed by the system (after servicing any page faults that may arise)? Show suitable calculations.

Ans: 60 = 0011 1100 = page 3. 3 causes page fault, replaces LRU page 0, and mapped to frame 2. So physical address = 0010 1100 = 44

33. Consider a 64-bit system running an OS that uses hierarchical page tables to manage virtual memory. Assume that logical and physical pages are of size 4KB and each page table entry is 4 bytes in size.

(a) What is the maximum number of levels in the page table of a process, including both the outermost page directory and the innermost page tables?

(b) Indicate which bits of the virtual address are used to index into each of the levels of the page table.

(c) Calculate the maximum number of pages that may be required to store all the page table entries of a process across all levels of the page table.

Ans

(a) ceil (64 - 12)/(12 - 2) = 6

(b) 2, 10, 10, 10, 10, 10 (starting from most significant to least)

(c) Innermost level has 2^{52} PTEs, which fit in 2^{42} pages. The next level has 2^{42} PTEs which require 2^{32} pages, and so on. Total pages = $2^{42} + 2^{32} + 2^{22} + 2^{12} + 2^2 + 1$

34. The page size in a system (running a Linux-like operating sytem on x86 hardware) is increased while keeping everything else (including the total size of main memory) the same. For each of the following metrics below, indicate whether the metric is *generally* expected to increase, decrease, or not change as a result of this increase in page size.

    (a) Size of the page table of a process

    (b) TLB hit rate

    (c) Internal fragmentation of main memory

    Ans: (a) PT size decreases (fewer entries) (b) TLB hit rate increases (more coverage) (c) Internal fragmentation increases (more space wasted in a page)

35. Consider a process with 4 logical pages, numbered 0–3. The page table of the process consists of the following logical page number to physical frame number mappings: (0, 11), (1, 35), (2, 3), (3, 1). The process runs on a system with 16 bit virtual addresses and a page size of 256 bytes. You are given that this process accesses virtual address 770. Answer the following questions, showing suitable calculations.

    (a) Which logical page number does this virtual address correspond to?

    (b) Which physical address does this virtual address translate to?

    Ans: (a) 770 = 512 + 256 + 2 = 00000011 00000010 = page 3, offset 2

    (b) page 3 maps to frame 1. physical address = 0000001 00000010 = 256 + 2 = 258

36. Consider a system with 16 bit virtual addresses, 256 byte pages, and 4 byte page table entries. The OS builds a multi-level page table for each process. Calculate the maximum number of pages required to store all levels of the page table of a process in this system.

    Ans: Number of PTE per process = $2^{16}/2^8 = 2^8$. Number of PTE per page = $2^8/2^2 = 2^6$. Number of inner page table pages = $2^8/2^6$ = 4, which requires one outer page directory. So total pages = 4+1 = 5.

37. Consider a process with 4 physical pages numbered 0–3. The process accesses pages in the following sequence: 0, 1, 0, 2, 3, 3, 0, 2. Assume that the RAM can hold only 3 out of these 4 pages, is initially empty, and there is no other process executing on the system.

    (a) Assuming the demand paging system is using an LRU replacement policy, how many page faults do the 8 page accesses above generate? Indicate the accesses which cause the faults.

    (b) What is the minimum number of page faults that would be generated by an optimal page replacement policy? Indicate the accesses which cause the faults.

Ans:

(a) 0 (M), 1 (M), 0(H), 2 (M), 3 (M), 3(H), 0(H), 2(H) = 4 misses

(b) Same as above

38. Consider a Linux-like operating system running on a 48-bit CPU hardware. The OS uses hierarchical paging, with 8 KB pages and 4 byte page table entries.

    (a) What is the maximum number of levels in the page table of a process, including both the outermost page directory and the innermost page tables?

    Ans: ceil (48 - 13)/(13 - 2) = 4

    (b) Indicate which bits of the virtual address are used to index into each of the levels of the page table.

    Ans: 2, 11, 11, 11

    (c) Calculate the maximum number of pages that may be required to store all the page table entries of a process across all levels of the page table.

    Ans: Innermost level has 2^{35} PTEs. Each page can accommodate 2^{11} PTEs. Total pages = $2^{24} + 2^{13} + 2^2 + 1$

39. Consider the scenario described in the previous question. You are told that the OS uses demand paging. That is, the OS allocates a physical frame and a corresponding PTE in the page table only when the memory location is accessed for the first time by a process. Further, the pages at all levels of the hierarchical page table are also allocated on demand, i.e., when there is at least one valid PTE within that page. A process in this system has accessed memory locations in 4K unique pages so far. You may assume that none of these 4K pages has been swapped out yet. You are required to compute the minimum and maximum possible sizes of the page table of this process after all accesses have completed.

    (a) What is the minimum possible size (in pages) of the page table of this process?

    Ans: Each page holds 2^{11} PTEs. So 2^{12} pages can be accommodated in 2 pages at the inner most level. Minimum pages in each of the outer levels is 1. So minimum size = 2 + 1 + 1 + 1 = 5.

    (b) What is the maximum possible size (in pages) of the page table of this process?

    Ans: The 2^{12} pages/PTEs could have been widely spread apart and in distinct pages at all levels of the page table. So maximum size = $2^{12} + 2^{12} + 2^2 + 1$.

40. In a demand paging system, it is intuitively expected that increasing the number of physical frames will naturally lead to a reduction in the rate of page faults. However, this intuition does not hold for some page replacement policies. A replacement policy is said to suffer from *Belady's anomaly* if increasing the number of physical frames in the system can sometimes lead to an increase in the number of page faults. Consider two page replacement policies studied in class: FIFO and LRU. For each of these two policies, you must state if the policy can suffer from Belady's anomaly (yes/no). Further, if you answer yes, you must provide an example of the occurrence of the anomaly, where increasing the number of physical frames actually leads to an increase in the number of page faults. If you answer no, you must provide an explanation of why you think the anomaly can never occur with this policy.

Hint: you may consider the following example. A process has 5 logical pages, and accesses them in this order: 1, 2, 3, 4, 1, 2, 5, 1, 2, 3, 4, 5. You may find this scenario useful in finding an example of Belady's anomaly. Of course, you may use any other example as well.

(a) FIFO

Ans: Yes. For string above, 9 faults with 3 frames and 10 faults with 4 frames.

(b) LRU

Ans: No. The N most recently used frames are always a subset of N+1 most recently used frames. So if a page fault occurs with N+1 frames, it must have occurred with N frames also. So page faults with N+1 frames can never be higher.

Lectures on Operating Systems (Mythili Vutukuru, IIT Bombay)

Problem Set: Networking

1. Consider a high performance networking application running on a high end multicore server. It is found that, under high incoming packet load, the system spends a large fraction of its time handling interrupts and context switches, leading to very little productive work at the application layer. Suggest one mechanism by which this problem can be mitigated. (For this question and the next, you are required to provide a description of the mechanism, not just its name.)

   Ans: Polling or batching interrupts

2. The current Linux network stack copies packet buffers several times, from the device to kernel space to user space. Suggest one mechanism by which the overhead of packet copies can be minimized, in order to build a high performance network stack.

   Ans: Directly DMA into user space (DPDK), or mmap kernel packet buffers to userspace (netmap).

3. Consider a web server that uses non-blocking event-driven I/O for network communication, but uses blocking I/O to access the disk. The web server wishes to run as multiple processes, so that the server can be available even if some subset of the processes block on disk I/O. Further, the web server wishes to receive web requests only on port 80, and not at different ports in the different processes. Suggest one mechanism by which the multiple server processes can handle requests arriving on a single port on the system.

   Ans: The master process can open a socket on port 80, fork multiple processes, and all child processes can accept connections from the same socket on port 80 using locks for mutual exclusion. Or, the master process can alone listen to requests on port 80 and assigns all blocking disk I/O to worker processes via IPC mechanisms.

4. Below are several problems with the kernel network stack that arise in multicore systems desiring high-performance network I/O. For each problem below, describe one technique that attempts to solve the stated problem. You are required to provide a 1–2 sentence description of the mechanism and how it fixes the stated problem, and not just its name.

   (a) Buffers to store packets are dynamically allocated and deallocated in the kernel, leading to dynamic memory allocation overheads.

   (b) The payload of a packet is copied multiple times, once from the NIC to kernel memory, and once again from kernel memory to userspace memory.

   (c) Multiple threads of an application running on different cores all contend for a lock to accept connections on the shared listen socket.

   Ans:

(a) Preallocate a circular ring of packet buffers and expose them to userspace processes (netmap and DPDK).

(b) Memory map the NIC ring into user space (netmap) or provide a userspace packet buffer ring to the NIC via a userspace device driver (DPDK).

(c) Have per-core accept queues.

5. Consider a multicore system running a TCP-based multi-threaded key-value store application. The incoming traffic to the system consists of new TCP connection requests, and get/put requests over the established TCP connections. In order to distribute the interrupt load across all cores, the NIC partitions incoming packets into multiple hardware queues using the hash of the 4-tuple (source and destination IP address and port) of the packet. The interrupts from each hardware queue are delivered to a separate core. The interrupts are processed via the regular Linux network stack on the various cores thereafter. The key-value store application consists of multiple threads, all of which access a shared hashmap data structure containing the key-value pairs.

   (a) Are the interrupts generated for all packets of a certain TCP flow guaranteed to be delivered to the same core by the NIC? Answer yes/no and justify.

   (b) Are all `get` requests for a certain key guaranteed to be handled by the same core at the application layer? Answer yes/no and justify.

   Ans:

   (a) Yes, because all packets of a flow will have the same 4-tuple hash.

   (b) No, because a request of a key can be sent over different TCP connections with different 4 tuple hashes, and hence can be processed by different cores.

6. Consider a TCP server socket program written in an event-driven manner. The server receives requests from multiple concurrent clients. The main thread of the server monitors read events on the server listening socket as well as all client sockets using the `select` or `epoll` family of system calls. When not processing any events, the server always blocks in an event-driven wait loop, e.g., `epoll_wait`, waiting for notifications. To process a client request, the server must read some data from the disk and send a reply back to the client. In order to avoid blocking the main event-driven server thread during disk reads, the server uses worker threads to block on disk I/O. After reading a client request, the main server thread spawns a new worker thread and passes the client request to the thread. This worker initiates the disk read and blocks until the disk read completes, while the main server thread continues to process other events on the sockets. Once a worker thread completes the disk read, it places the data read from the disk in a shared datastructure that is accessible by the main server thread (using suitable locking), and terminates. Assume that the main server thread does not automatically get any notification from the OS when the worker thread terminates. Now, we require that the server send a response back to the client once the worker thread has completed the disk read operation. There are multiple mechanisms to accomplish this goal, and the rest of the question lets you explore the various design choices.

   (a) One way in which the server can send responses back to the clients is by monitoring the status of the shared datastructure every time an event occurs, i.e., when the main server thread returns from the `select` or `epoll_wait` system calls. At this time, if the server finds that some worker has placed a response in the shared buffer, it can fetch this response

and send it back to the client on its socket. This solution is almost but not fully correct. Describe one scenario where this solution will fail.

(b) Describe how you can fix the server design to overcome the above failure scenario.

Ans: (a) if only one request and no other traffic comes in, server will never wake up from epoll loop (b) Have a unix socket from worker to main thread. When worker finishes, it will write to the unix socket, which will cause the server to wake up from the event loop.

7. This question will explore the design of a simple multi-process file server. The server is expected to serve multiple concurrent clients. Each client opens a TCP connection to the server, and sends requests for files on that connection. The web server must read the name of the requested file from the socket, read the file from disk, write it back into the socket, and wait for the client to request the next file. The server must serve the client in this manner until the client closes the connection.

   You are given a single-threaded file server running as a single process on a multicore system with a Linux-like OS. Because the server performs multiple blocking operations (accepting new connections, reading from sockets, reading from disk), a single server process can neither effectively serve multiple clients nor efficiently utilize the multiple CPU cores. In order to overcome these problems, you must modify the server to fork some children and delegate work to the child processes. Note that you are constrained to increase the parallelism of the server only by spawning processes and not threads. Further, you may not make any of the blocking operations (e.g., socket reads) non-blocking.

   The following sub-parts of the question will let you describe the design of your multiprocess file server. Note that multiple correct designs are possible; it is sufficient if you describe one of the possible correct designs.

   (a) When (i.e., at what point in the server's code) does the main server process spawn a child process? And when does it reap this process?

   (b) How is the work of handling a client (accepting a connection, reading from socket, reading from disk, writing to socket) divided between the parent and child processes? In other words, who does what part of the work?

   (c) How do the parent and child processes exchange information with each other? Note that the communication can be implicit via variables or file descriptors inherited at the time of fork, or explicit via IPC mechanisms You must describe all such exchange of information between the parent and child processes.

   (d) State any other aspect of your design not covered by the questions above, and any assumptions you have made in your solution.

   Ans: One possible design is spawn a child every time a new client connects and accept returns. The parent process periodically (say before calling accept) reaps any dead children. The child process does the network read, disk read, network write for the client, while the main process does the job of accepting new connections. The child gets the new file descriptor of the client to serve via a variable implicitly from the parent.

8. Which **one** of the following system calls initiates the three-way TCP handshake?

   (a) `socket`

(b) `listen`

(c) `connect`

(d) `accept`

Ans: c

9. A modern DMA-capable NIC has received a packet over the network, and has raised an interrupt. A modern Linux-like OS executes an interrupt handler to service this interrupt. Which **one** of the following operations is executed by the OS during the top-half of the interrupt handler?

   (a) Copy the packet buffer from NIC memory to kernel memory

   (b) Copy the packet buffer from kernel memory to user memory

   (c) Updates to the RX ring pointers

   (d) Generation of TCP acknowledgement for the received packet

   Ans: c

10. Consider a web server system consisting of N replicas. Clients send HTTP requests for web objects (HTML pages, images, etc.) to the server over TCP connections. All requests are sent to a single server IP address and port that is publicized to clients (say, via DNS). A load balancer placed before the replicas rewrites the destination IP address of the packets coming to the server to redistribute traffic to the various server replicas. For every packet arriving for the web server, the load balancer computes the hash h of the TCP/IP header 4-tuple (source/destination IP address/port), computes $i = h \bmod N$, and redirects the packet to the i-th server replica. Ignore any changes to the set of replicas, or any failures.

    (a) Are all packets of a given TCP connection always sent to the same replica? Answer yes/no, and justify.

    (b) Are all requests for a given HTTP web object always sent to the same replica? Answer yes/no, and justify.

    Ans: (a) yes because all packets of a connection hash to same replica (b) no because requests of a web object can come over multiple connections, and can hash to different replicas

Lectures on Operating Systems (Mythili Vutukuru, IIT Bombay)

Practice Problems: Processes

1. Answer yes/no, and provide a brief explanation.

   (a) Can two processes be concurrently executing the same program executable?

   (b) Can two running processes share the complete process image in physical memory (not just parts of it)?

   Ans:

   (a) Yes, two processes can run the same program.

   (b) No, in general. (Only time this is possible is with copy-on-write during fork, and before any writes have been made.)

2. Consider a process executing on a CPU. Give an example scenario that can cause the process to undergo:

   (a) A voluntary context switch.

   (b) An involuntary context switch.

   Ans:

   (a) A blocking system call.

   (b) Timer interrupt that causes the process to be switched out.

3. Consider a parent process P that has forked a child process C. Now, P terminates while C is still running. Answer yes/no, and provide a brief explanation.

   (a) Will C immediately become a zombie?

   (b) Will P immediately become a zombie, until reaped by its parent?

   Ans:

   (a) No, it will be adopted by init.

   (b) Yes.

4. A process in user mode cannot execute certain privileged hardware instructions. [T/F]

   Ans: True, some instructions in every CPU's instruction set architecture can only be executed when the CPU is running in a privileged mode (e.g., ring 0 on Intel CPUs).

5. Which of the following C library functions do NOT directly correspond to (similarly named) system calls? That is, the implentations of which of these C library functions are NOT straightforward invocations of the underlying system call?

   (a) `system`, which executes a bash shell command.

   (b) `fork`, which creates a new child process.

   (c) `exit`, which terminates the current process.

   (d) `strlen`, which returns the length of a string.

   Ans: (a), (d)

6. Which of the following actions by a running process will *always* result in a context switch of the running process, even in a non-preemptive kernel design?

   (a) Servicing a disk interrupt, that results in another blocked process being marked as ready/runnable.

   (b) A blocking system call.

   (c) The system call exit, to terminate the current process.

   (d) Servicing a timer interrupt.

   Ans: (b), (c)

7. Consider two machines A and B of different architectures, running two different operating systems OS-A and OS-B. Both operating systems are POSIX compliant. The source code of an application that is written to run on machine A must always be rewritten to run on machine B. [T/F]

   Ans: False. If the code is written using the POSIX API, it need not be rewritten for another POSIX compliant system.

8. Consider the scenario of the previous question. An application binary that has been compiled for machine A may have to be recompiled to execute correctly on machine B. [T/F]

   Ans: True. Even if the code is POSIX compliant, the CPU instructions in the compiled executable are different across different CPU architectures.

9. A process makes a system call to read a packet from the network device, and blocks. The scheduler then context-switches this process out. Is this an example of a voluntary context switch or an involuntary context switch?

   Ans: Voluntary context switch.

10. A context switch can occur only after processing a timer interrupt, but not after any other system call or interrupt. [T/F]

    Ans: False, a context switch can also occur after a blocking system call for example.

11. A C program cannot directly invoke the OS system calls and must always use the C library for this purpose. [T/F]

    Ans: False, it is cumbersome but possible to directly invoke system calls from user code.

12. A process undergoes a context switch every time it enters kernel mode from user mode. [T/F]

    Ans: False, after finishing its job in kernel mode, the OS may sometimes decide to go back to the user mode of the same process, without switching to another process.

13. Consider a process P in xv6 that invokes the wait system call. Which of the following statements is/are true?

    (a) If P does not have any zombie children, then the wait system call returns immediately.

    (b) The wait system call always blocks process P and leads to a context switch.

    (c) If P has exactly one child process, and that child has not yet terminated, then the wait system call will cause process P to block.

    (d) If P has two or more zombie children, then the wait system call reaps all the zombie children of P and returns immediately.

    Ans: (c)

14. Consider a process P that executes the fork system call twice. That is, it runs code like this:

    int ret1 = fork(); int ret2 = fork();

    How many direct children of P (i.e., processes whose parent is P) and how many other descendants of P (i.e., processes who are not direct children of P, but whose grandparent or great grandparent or some such ancestor is P) are created by the above lines of code? You may assume that all fork system calls succeed.

    (a) Two direct children of P are created.

    (b) Four direct children of P are created.

    (c) No other descendant of P is created.

    (d) One other descendant of P is created.

    Ans: (a), (d)

15. Consider the x86 instruction"int n" that is executed by the CPU to handle a trap. Which of the following statements is/are true?

    (a) This instruction is always invoked by privileged OS code.

    (b) This instruction causes the CPU to set its EIP to address value "n".

    (c) This instruction causes the CPU to lookup the Interrupt Descriptor Table (IDT) using the value "n" as an index.

    (d) This instruction is always executed by the CPU only in response to interrupts from external hardware, and never due to any code executed by the user.

    Ans: (c)

16. Consider the following scheduling policy implemented by an OS, in which a user can set numerical priorities for processes running in the system. The OS scheduler maintains all ready processes in a strict priority queue. When the CPU is free, it extracts the ready process with the highest priority (breaking ties arbitrarily), and runs it until the process blocks or terminates. Which of the following statements is/are true about this scheduling policy?

(a) This scheduler is an example of a non-preemptive scheduling policy.

(b) This scheduling policy can result in the starvation of low priority processes.

(c) This scheduling policy guarantees fairness across all active processes.

(d) This scheduling policy guarantees lowest average turnaround time for all processes.

Ans: (a), (b)

17. Consider the following scheduling policy implemented by an OS. Every time a process is scheduled, the OS runs the process for a maximum of 10 milliseconds or until the process blocks or terminates itself before 10 milliseconds. Subsequently, the OS moves on to the next ready process in the list of processes in a round-robin fashion. Which of the following statements is/are true about this scheduling policy?

    (a) This policy cannot be efficiently implemented without hardware support for timer interrupts.

    (b) This scheduler is an example of a non-preemptive scheduling policy.

    (c) This scheduling policy can sometimes result in involuntary context switches.

    (d) This scheduling policy prioritizes processes with shorter CPU burst times over processes that run for long durations.

    Ans: (a), (c)

18. Consider a process P that needs to save its CPU execution context (values of some CPU registers) on some stack when it makes a function call or system call. Which of the following statements is/are true?

    (a) During a system call, when transitioning from user mode to kernel mode, the context of the process is saved on its kernel stack.

    (b) During a function call in user mode, the context of the process is saved on its user stack.

    (c) During a function call in kernel mode, the context of the process is saved on its user stack.

    (d) During a function call in kernel mode, the context of the process is saved on its kernel stack.

    Ans: (a), (b), (d)

19. Which of the following statements is/are true regarding how the trap instruction (e.g., int n in x86) is invoked when a trap occurs in a system?

    (a) When a user makes a system call, the trap instruction is invoked by the kernel code handling the system call

    (b) When a user makes a system call, the trap instruction is invoked by userspace code (e.g., user program or a library)

    (c) When an external I/O device raises an interrupt, the trap instruction is invoked by the device driver handling the interrupt

    (d) When an external I/O device raises an interrupt signaling the completion of an I/O request, the trap instruction is invoked by the user process that raised the I/O request

    Ans: (b)

20. Which of the following statements is/are true about a context switch?

    (a) A context switch from one process to another will happen every time a process moves from user mode to kernel mode

    (b) For preemptive schedulers, a trap of any kind always leads to a context switch

    (c) A context switch will always occur when a process has made a blocking system call, irrespective of whether the scheduler is preemptive or not

    (d) For non-preemptive schedulers, a process that is ready/willing to run will not be context switched out

    Ans: (c), (d)

21. Consider the following C program. Assume there are no syntax errors and the program executes correctly. Assume the fork system calls succeed. What is the output printed to the screen when we execute the below program?

```
void main(argc, argv) {

  for(int i = 0; i < 4; i++) {
    int ret = fork();
    if(ret == 0)
      printf("child %d\n", i);
  }
}
```

    Ans: The statement "child i" is printed 2^i times for i=0 to 3

22. Consider a parent process P that has forked a child process C in the program below.

```
int a = 5;
int fd = open(...) //opening a file
int ret = fork();
if(ret >0) {
  close(fd);
  a = 6;
  ...
}
else if(ret==0) {
  printf("a=%d\n", a);
  read(fd, something);
}
```

    After the new process is forked, suppose that the parent process is scheduled first, before the child process. Once the parent resumes after fork, it closes the file descriptor and changes the value of a variable as shown above. Assume that the child process is scheduled for the first time only after the parent completes these two changes.

(a) What is the value of the variable `a` as printed in the child process, when it is scheduled next? Explain.

(b) Will the attempt to read from the file descriptor succeed in the child? Explain.

Ans:

(a) 5. The value is only changed in the parent.

(b) Yes, the file is only closed in the parent.

23. Consider the following pseudocode. Assume all system calls succeed and there are no other errors in the code.

```
int ret1 = fork(); //fork1
int ret2 = fork(); //fork2
int ret3 = fork(); //fork3
wait();
wait();
wait();
```

Let us call the original parent process in this program as P. Draw/describe a family tree of P and all its descendents (children, grand children, and so on) that are spawned during the execution of this program. Your tree should be rooted at P. Show the spawned descendents as nodes in the tree, and connect processes related by the parent-child relationship with an arrow from parent to child. Give names of the form C¡number¿ for descendents, where child processes created by fork "i" above should have numbers like "i1", "i2", and so on. For example, child processes created by fork3 above should have names C31, C32, and so on.

Ans: P has three children, one in each fork statement: C11, C21, C31. C11 has two children in the second and third fork statements: C22, C32. C21 and C22 also have a child each in the third fork statement: C33 and C34.

24. Consider a parent process that has forked a child in the code snippet below.

```
int count = 0;
ret = fork();
if(ret == 0)  {
printf("count in child=%d\n", count);
}
else  {
count = 1;
}
```

The parent executes the statement "count = 1" before the child executes for the first time. Now, what is the value of count printed by the code above? Assume that the OS implements a simple fork (not a copy-on-write fork).

Ans: 0 (the child has its own copy of the variable)

25. Consider the wait family of system calls (wait, waitpid etc.) provided by Linux. A parent process uses some variant of the wait system call to wait for a child that it has forked. Which of the following statements is always true when the parent invokes the system call?

    (a) The parent will always block.

    (b) The parent will never block.

    (c) The parent will always block if the child is still running.

    (d) Whether the parent will block or not will depend on the system call variant and the options with which it is invoked.

    Ans: (d)

26. Consider a simple linux shell implementing the command `sleep 100`. Which of the following is an accurate ordered list of system calls invoked by the shell from the time the user enters this command to the time the shell comes back and asks the user for the next input?

    (a) wait-exec-fork

    (b) exec-wait-fork

    (c) fork-exec-wait

    (d) wait-fork-exec

    Ans: (c)

27. Consider a process P1 that forks P2, P2 forks P3, and P3 forks P4. P1 and P2 continue to execute while P3 terminates. Now, when P4 terminates, which process must wait for and reap P4?

    Ans: init (orphan processes are reaped by init)

28. Consider the following three processes that arrive in a system at the specified times, along with the duration of their CPU bursts. Process P1 arrives at time t=0, and has a CPU burst of 10 time units. P2 arrives at t=2, and has a CPU burst of 2 units. P3 arrives at t=3, and has a CPU burst of 3 units. Assume that the processes execute only once for the duration of their CPU burst, and terminate immediately. Calculate the time of completion of the three processes under each of the following scheduling policies. For each policy, you must state the completion time of all three processes, P1, P2, and P3. Assume there are no other processes in the scheduler's queue. For the preemptive policies, assume that a running process can be immediately preempted as soon as the new process arrives (if the policy should decide to premempt).

    (a) First Come First Serve

    (b) Shortest Job First (non-preemptive)

    (c) Shortest Remaining Time First (preemptive)

    (d) Round robin (preemptive) with a time slice of (atmost) 5 units per process

    Ans:

    (a) FCFS: P1 at 10, P2 at 12, P3 at 15

    (b) SJF: same as above

(c) SRTF: P2 at 4, P3 at 7, P1 at 15

(d) RR: P2 at 7, P3 at 10, P1 at 15

29. Consider an application that is composed of one master process and multiple worker processes that are forked off the master at the start of application execution. All processes have access to a pool of shared memory pages, and have permissions to read and write from it. This shared memory region (also called the request buffer) is used as follows: the master process receives incoming requests from clients over the network, and writes the requests into the shared request buffer. The worker processes must read the request from the request buffer, process it, and write the response back into the same region of the buffer. Once the response has been generated, the server must reply back to the client. The server and worker processes are single-threaded, and the server uses event-driven I/O to communicate over the network with the clients (you must not make these processes multi threaded). You may assume that the request and the response are of the same size, and multiple such requests or responses can be accommodated in the request buffer. You may also assume that processing every request takes similar amount of CPU time at the worker threads.

    Using this design idea as a starting point, describe the communication and synchronization mechanisms that must be used between the server and worker processes, in order to let the server correctly delegate requests and obtain responses from the worker processes. Your design must ensure that every request placed in the request buffer is processed by one and only one worker thread. You must also ensure that the system is efficient (e.g., no request should be kept waiting if some worker is free) and fair (e.g., all workers share the load almost equally). While you can use any IPC mechanism of your choice, ensure that your system design is practical enough to be implementable in a modern multicore system running an OS like Linux. You need not write any code, and a clear, concise and precise description in English should suffice.

    Ans: Several possible solutions exist. The main thing to keep in mind is that the server should be able to assign a certain request in the buffer to a worker, and the worker must be able to notify completion. For example, the master can use pipes or sockets or message queues with each worker. When it places a request in the shared memory, it can send the position of the request to one of the workers. Workers listen for this signal from the master, process the request, write the response, and send a message back to the master that it is done. The master monitors the pipes/sockets of all workers, and assigns the next request once the previous one is done.

30. Consider the following events that happen during a context switch from (user mode of) process P to (user mode of) process Q, triggered by a timer interrupt that occurred when P was executing, in a Unix-like operating system design studied in class. Arrange the events in chronological order, starting from the earliest to the latest.

    (A) The CPU program counter moves from the kernel address space of P to the kernel address space of Q.

    (B) The CPU executing process P moves from user mode to kernel mode.

    (C) The CPU stack pointer moves from the kernel stack of P to the kernel stack of Q.

    (D) The CPU program counter moves from the kernel address space of Q to the user address space of Q.

    (E) The OS scheduler code is invoked.

    Ans:

B E C A D

31. Consider a system with two processes P and Q, running a Unix-like operating system as studied in class. Consider the following events that may happen when the OS is concurrently executing P and Q, while also handling interrupts.

    (A) The CPU program counter moves from pointing to kernel code in the kernel mode of process P to kernel code in the kernel mode of process Q.

    (B) The CPU stack pointer moves from the kernel stack of P to the kernel stack of Q.

    (C) The CPU executing process P moves from user mode of P to kernel mode of P.

    (D) The CPU executing process P moves from kernel mode of P to user mode of P.

    (E) The CPU executing process Q moves from the kernel mode of Q to the user mode of Q.

    (F) The interrupt handling code of the OS is invoked.

    (G) The OS scheduler code is invoked.

    For each of the two scenarios below, list out the chronological order in which the events above occur. Note that all events need not occur in each question.

    (a) A timer interrupt occurs when P is executing. After processing the interrupt, the OS scheduler decides to return to process P.

    (b) A timer interrupt occurs when P is executing. After processing the interrupt, the OS scheduler decides to context switch to process Q, and the system ends up in the user mode of Q.

    Ans:

    (a) C F G D

    (b) C F G B A E

32. Which of the following pieces of information in the PCB of a process are changed when the process invokes the exec system call?

    (a) Process identifier (PID)

    (b) Page table entries

    (c) The value of the program counter stored within the user space context on the kernel stack

    Ans: (a) does not change. (b) and (c) change because the process gets a new memory image (and hence new page table entries pointing to the new image).

33. Which of the following pieces of information about the process are identical for a parent and the newly created child processes, immediately after the completion of the `fork` system call? Answer "identical" or "not identical".

    (a) The process identifier.

    (b) The contents of the file descriptor table.

    Ans: (a) is not identical, as every process has its own unique PID in the system. (b) is identical, as the child gets an exact copy of the parent's file descriptor table.

34. Consider the following sample code from a simple shell program.

```
command = read_from_user();
int rc = fork();
if(rc == 0) { //child
  exec(command);
}
else {//parent
  wait();
}
```

Now, suppose the shell wishes the redirect the output of the command not to STDOUT but to a file "foo.txt". Show how you would modify the above code to achieve this output redirection. You can indicate your changes next to the code above.

Ans:

Modify the child code as follows.

```
close(STDOUT_FILENO)
open("foo.txt")
exec(command)
```

35. Consider a simple program shown below. The OS uses a copy-on-write fork implementation. Indicate the line of code whose execution causes the OS to start making two separate copies of the memory image for the parent and child processes. Assume that the parent process is scheduled to run before the child after the fork system call.

```
int a = 0;
int rc = fork();
if(rc == 0) { //child
  a = -1;
  exec(some_other_executable);
}
else {//parent
  a = 1;
  wait();
}
```

Ans: The line a = 1 in parent.

36. What is the output of the following code snippet? You are given that the `exec` system call in the child does not succeed.

```
int ret = fork();
if(ret==0) {
  exec(some_binary_that_does_not_exec);
  printf(``child\n'');
```

```
  }
else {
  wait();
  printf(``parent\n'');
  }
```

Ans:

```
child
parent
```

37. When a process makes a system call and runs kernel code:

    (a) How does the process obtain the address of the kernel instruction to jump to?

    (b) Where is the userspace context of the process (program counter and other registers) stored during the transition from user mode to kernel mode?

    Ans:

    (a) From IDT (interrupt descriptor table) (b) on kernel stack of process (which is linked from the PCB)

38. Consider a process P1 that is executing on a Linux-like OS on a single core system. When P1 is executing, a disk interrupt occurs, causing P1 to go to kernel mode to service that interrupt. The interrupt delivers all the disk blocks that unblock a process P2 (which blocked earlier on the disk read). The interrupt service routine has completed execution fully, and the OS is just about to return back to the user mode of P1. At this point in time, what are the states (ready/running/blocked) of processes P1 and P2?

    (a) State of P1

    (b) State of P2

    Ans:

    (a) P1 is running (b) P2 is ready

39. Consider the following code snippet, where a parent process forks a child process. The child performs one task during its lifetime, while the parent performs two different tasks.

    ```
    int ret = fork();
    if(ret == 0) { do_child_task(); }
    else { do_parent_task1();
           do_parent_task2(); }
    ```

    With the way the code is written right now, the user has no control over the order in which the parent and child tasks execute, because the scheduling of the processes is done by the OS. Below are given two possible orderings of the tasks that the user wishes to enforce. For each part, briefly describe how you will modify the code given above to ensure the required ordering of tasks. You may write your answer in English or using pseudocode.

Note that you cannot change the OS scheduling mechanism in any way to solve this question. If a process is scheduled by the OS before you want its task to execute, you must use mechanisms like system calls and IPC techniques available to you in userspace to delay the execution of the task till a suitable time.

(a) We want the parent to start execution of both its tasks only after the child process has finished its task and has terminated.

Ans: Parent does wait() until child finishes, and then starts its tasks.

(b) We want the child process to execute its task after the parent process has finished its first task, but before it runs its second task. The parent must not execute its second task until the child has completed its task and has terminated.

Ans: Many solutions are possible. Parent and child share two pipes (or a socket). Parent writes to one pipe after completing task 1 and child blocks on this pipe read before starting its task. Child writes to pipe 2 after finishing its task, and parent blocks on this pipe read before starting its second task. (Or parent can use wait to block for child termination, like in previous part.)

40. Consider a system with a single CPU core and three processes A, B, C. Process A arrives at $t = 0$, and runs on the CPU for 10 time units before it finishes. Process B arrives at $t = 6$, and requires an initial CPU time of 3 units, after which it blocks to perform I/O for 3 time units. After returning from I/O wait, it executes for a further 5 units before terminating. Process C arrives at $t = 8$, and runs for 2 units of time on the CPU before terminating. For each of the scheduling policies below, calculate the time of completion of each of the three processes. Recall that only the size of the current CPU burst (excluding the time spent for waiting on I/O) is considered as the "job size" in these schedulers.

(a) First Come First Serve (non-preemptive).

Ans: A=10, B=21, C=15. A finishes at 10 units. First run of B finishes at 13. C completes at 15. B restarts at 16 and finishes at 21.

(b) Shortest Job First (non-preemptive)

Ans: A=10, B=23, C=12. A finishes at 10 units. Note that the arrival of shorter jobs B and C does not preempt A. Next, C finishes at 12. First task of B finishes at 15, B blocks from 15 to 18, and finally completes at 23 units.

(c) Shortest Remaining Time First (preemptive)

Ans: A=15, B=20, C=11. A runs until 6 units. Then the first task of B runs until 9 units. Note that the arrival of C does not preempt B because it has a shorter remaining time. C completes at 11. B is not ready yet, so A runs for another 4 units and completes at 15. Note that the completion of B's I/O does not preempt A because A's remaining time is shorter. B finally restarts at 15 and completes at 20.

41. Consider the following code snippet running on a modern Linux operating systems (with a reasonable preemptive scheduling policy as studied in class). Assume that there are no other interfering processes in the system. Note that the executable "good_long_executable" runs for 100 seconds, prints the line "Hello from good executable" to screen, and terminates. On the other hand, the file "bad_executable" does not exist and will cause the `exec` system call to fail.

```
int ret1 = fork();
if(ret1 == 0) { //Child 1
  printf("Child 1 started\n");
  exec("good_long_executable");
  printf("Child 1 finished\n");
}
else { //Parent
  int ret2 == fork();
  if(ret2 == 0) { //Child 2
    sleep(10); //Sleeping allows child 1 to begin execution
    printf("Child 2 started\n");
    exec("bad_executable");
    printf("Child 2 finished\n");
  } //end of Child 2
  else { //Parent
    wait();
    printf("Child reaped\n");
    wait();
    printf("Parent finished\n");
  }
}
```

Write down the output of the above program.

Ans:

Child 1 started
Child 2 started
(Some error message from the wrong executable)
Child 2 finished
Child reaped
Hello from good executable
Parent finished

42. What are the possible outputs printed from this program shown below? You may assume that the program runs on a modern Linux-like OS. You may ignore any output generated from "some_executable". You must consider all possible scenarios of the system calls succeeding as well as failing. In your answer, clearly list down all the possible scenarios, and the output of the program in each of these scenarios.

```
int ret = fork();
if(ret == 0) {
  printf(``Hello1\n'');
  exec(``some_executable'');
  printf(``Hello2\n'');
} else if(ret > 0) {
  wait();
```

```
  printf(``Hello3\n'');
} else {
  printf(``Hello4\n'');
}
```

Ans: Case I: fork and exec succeed. Hello1, Hello3 are printed. Case II: fork succeeds but exec fails. Hello1, Hello2, Hello3 are printed. Case III: fork fails. Hello4 is printed.

43. Which of the following operations by a process will definitely cause the process to move from user mode to kernel mode? Answer yes (if a change in mode happens) or no.

    (a) A process invokes a function in a userspace library.
    Ans: no

    (b) A process invokes the `kill` system call to send a signal to another process.
    Ans: yes

44. Consider the following sample code from a simple shell program.

```
int rc1 = fork();
if(rc1 == 0) {
  exec(cmd1);
}
else {
  int rc2 = fork();
  if(rc2 == 0) {
  exec(cmd2);
  }
  else {
    wait();
    wait();
  }
}
```

    (a) In the code shown above, do the two commands `cmd1` and `cmd2` execute serially (one after the other) or in parallel? **Ans:** parallel.

    (b) Indicate how you would modify the code above to change the mode of execution from serial to parallel or vice versa. That is, if you answered "serial" in part (a), then you must change the code to execute the commands in parallel, and vice versa. Indicate your changes next to the code snippet above. **Ans:** move the first wait to before second fork.

45. What is the output printed by the following snippet of pseudocode? If you think there is more than one possible answer depending on the execution order of the processes, then you must list all possible outputs.

```
int fd[2];
pipe(fd);
int rc = fork();
```

```
if(rc == 0) { //child
  close(fd[1]);
  printf(``child1\n'');
  read(fd[0], bufc, bufc_size);
  printf(``child2\n'');
}
else {//parent
  close(fd[0]);
  printf(``parent1\n'');
  write(fd[1], bufp, bufp_size);
  wait();
  printf(``parent2\n'');
}
```

Ans: If child scheduled before parent: child1, parent1, child2, parent 2. If parent scheduled before child, parent1, child1, child2, parent2.

Lectures on Operating Systems (Mythili Vutukuru, IIT Bombay)

Practice Problems: xv6 Filesystem

1. Consider two active processes in xv6 that are connected by a pipe. Process W is writing to the pipe continuously, and process R is reading from the pipe. While these processes are alternately running on the single CPU of the machine, no other user process is scheduled on the CPU. Also assume that these processes always give up CPU due to blocking on a full/empty pipe buffer rather than due to yielding on a timer interrupt . List the sequence of events that occur from the time W starts execution, to the time the context is switched to R, to the time it comes back to W again. You must list all calls to the functions `sleep` and `wakeup`, and calls to `sched` to give up CPU; clearly state which process makes these calls from which function. Further, you must also list all instances of acquiring and releasing `ptable.lock`. Make your description of the execution sequence as clear and concise as possible.

   Ans:

   (a) W calls `wakeup` to mark R as ready.
   (b) W realizes the pipe buffer is full and calls `sleep`.
   (c) W acquires `ptable.lock`.
   (d) W gives up the CPU in `sched`.
   (e) The OS performs a context switch from W to R and R starts execution.
   (f) R releases `ptable.lock`.
   (g) R calls `wakeup` to mark W as ready.
   (h) R realizes the pipe buffer is full and calls `sleep`.
   (i) R gives up the CPU in `sched`.
   (j) The OS performs a context switch from R to W and W starts execution

2. State one advantage of the disk buffer cache layer in xv6 besides caching. Put another way, even if there was zero cache locality, the higher layers of the xv6 file system would still have to use the buffer cache: state one functionality of the disk buffer cache (besides caching) that is crucial for the higher layers.

   Ans: Synchronization - only one process at a time handles a disk block.

3. Consider two processes in xv6 that both wish to read a particular disk block, i.e., either process does not intend to modify the data in the block. The first process obtains a pointer to the struct buf using the function “bread”, but never causes the buffer to become dirty. Now, if the second process calls “bread” on the same block before the first process calls ”brelse”, will this second call to “bread” return immediately, or would it block? Briefly describe what xv6 does in this case, and justify the design choice.

Ans: Second call to bread would block. Buffer cache only allows access to one block at a time, since the buffer cache has no control on how the process may modify the data

4. Consider a process that calls the log_write function in xv6 to log a changed disk block. Does this function block the invoking process (i.e., cause the invoking process to sleep) until the changed block is written to disk? Answer Yes/No.

   Ans: No

5. Repeat the previous question for when a process calls the bwrite function to write a changed buffer cache block to disk. Answer Yes/No.

   Ans: Yes

6. When the buffer cache in xv6 runs out of slots in the cache in the bget function, it looks for a clean LRU block to evict, to make space for the new incoming block. What would break in xv6 if the buffer cache implementation also evicted dirty blocks (by directly writing them to their original location on disk using the bwrite function) to make space for new blocks?

   Ans: All writes must happen via logging for consistent updates to disk blocks during system calls. Writing dirty blocks to disk bypassing the log will break this property.

7. (a) Recall that buffer caches of operating systems come in two flavors when it comes to writing dirty blocks from the cache to the secondary storage disk: write through caches and write back caches. Consider the buffer cache implementation in xv6, specifically the `bwrite` function. Is this implementation an example of a write through cache or a write back cache? Explain your answer.

   (b) If the xv6 buffer cache implementation changed from one mode to the other, give an example of xv6 code that would break, and describe how you would fix it. In other words, if you answered "write through" to part (a) above, you must explain what would go wrong (and how you would fix it) if xv6 moved to a write back buffer cache implementation. And if you answered "write back" to part (a), explain what would need to change if the buffer cache was modified to be write through instead.

   (c) The buffer cache in xv6 maintains all the `struct buf` buffers in a fixed-size array. However, an additional linked list structure is imposed on these buffers. For example, each `struct buf` also has pointers `struct buf *prev` and `struct buf *next`. What additional functions do these pointers serve, given that the buffers can all be accessed via the array anyway?

   Ans:

   (a) Write through cache

   (b) If changed to write back, the logging mechanism would break.

   (c) Helps implement LRU eviction

8. Consider a system running xv6. A process has the three standard file descriptors (0,1,2) open and pointing to the console. All the other file descriptors in its `struct proc` file descriptor table are unused. Now the process runs the following snippet of code to open a file (that exists on disk, but has not been opened before), and does a few other things as shown below. Draw a figure showing the file descriptor table of the process, relevant entries in the global open file table, and

the in-memory inode structures pointed at by the file table, after each point in the code marked parts (a), (b), and (c) below. Your figures must be clear enough to understand what happens to the kernel data structures right after these lines execute. You must draw three separate figures for parts (a), (b), and (c).

```
int fd;
fd = open("foo.txt", O_RDWR); //part (a)
dup(fd);                      //part (b)
fd = open("foo.txt", O_RDWR); //part (c)
```

Ans:

(a) Open creates new FD, open file table entry, and allocated new inode.

(b) Dup creates new FD to point to same open file table entry.

(c) Next open creates new open file table entry to point to the same inode.

9. Consider the execution of the system call `open` in xv6, to create and open a new file that does not already exist.

   (a) Describe all the changes to the disk and memory data structures that happen during the process of creating and opening the file. Write your answer as a bulleted list; each item of the list must describe one data structure, specify whether it is in disk or memory, and briefly describe the change that is made to this data structure by the end of the execution of the open system call.

   (b) Suggest a suitable ordering of the changes above that is most resilient to inconsistencies caused by system crashes. Note that the ordering is not important in xv6 due to the presence of a logging mechanism, but you must suggest an order that makes sense for operating systems without such logging features.

   Ans:

   (a)
   - A free inode on disk is marked as allocated for this file.
   - An inode from the in-memory inode cache is allocated to hold data for this new inode number.
   - A directory entry is written into the parent directory on disk, to point to the new inode.
   - An entry is created in the in-memory open file table to point to the inode in cache.
   - An entry is added to the in-memory per-process file descriptor table to point to the open file table entry.

   (b) The directory entry should be added after the on-disk inode is marked as free. The memory operations can happen in any order, as the memory data structures will not survive a crash.

10. Consider the operation of adding a (hard) link to an existing file /D1/F1 from another location /D2/F2 in the xv6 OS. That is, the linking process should ensure that accessing /D2/F2 is equivalent to accessing /D1/F1 on the system. Assume that the contents of all directories and files fit within one data block each. Let $i(x)$ denote the block number of the inode of a file/directory, and let $d(x)$ denote the block number of the (only) data block of a file/directory. Let L denote the starting

block number of the log. Block L itself holds the log header, while blocks starting $L + 1$ onwards hold data blocks logged to the disk.

Assume that the buffer cache is initially empty, except for the inode and data (directory entries) of the root directory. Assume that no other file system calls are happening concurrently. Assume that a transaction is started at the beginning of the link system call, and commits right after the end of it. Make any other reasonable asusmptions you need to, and list them down.

Now, list and explain all the read/write operations that the disk (not the buffer cache) sees during the execution of this link operation. Write your answer as a bulleted list. Each bullet must specify whether the operation is a read or write, the block number of the disk request, and a brief explanation on why this request to disk happens. Your answer must span the entire time period from the start of the system call to the end of the log commit process.

Ans:

- read i(D1), read d(D1)—this will give us the inode number of F1.
- read i(F1). After reading the inode and bringing it to cache, its link count will be updated. At this point, the inode is only updated in the buffer cache.
- read i(D2), read d(D2)—we check that the new file name F2 does not exist in the directory. After this, the directory contents are updated, and a note is made in the log.
- Now, the log starts committing. This transaction has two modified blocks (the inode of F1 and the directory content of D2). So we will see two disk blocks written to the log: write to L+1 and L+2, followed by a write to block L (the log header).
- Next, the transactions are installed: a write to disk blocks i(F1) and d(D2).
- Finally, another write to block L to clear the transaction header.

11. Which of the following statements is/are true regarding the file descriptor (FD) layer in xv6?

A. The FD returned by `open` is an index to the global `struct ftable`.

B. The FD returned by `open` is an index to the open file table that is part of the `struct proc` of the process invoking `open`.

C. Each entry in the global `struct ftable` can point to either an in-memory inode object or a pipe object, but never to both.

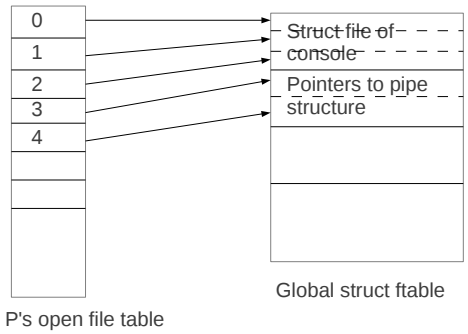
D. The reference count stored in an entry of the `struct ftable` indicates the number of links to the file in the directory tree.

Ans: BC

12. Consider the following snippet of the shell code from xv6 that implements pipes.

```
pcmd = (struct pipecmd*)cmd;
if(pipe(p) < 0)
  panic("pipe");
if(fork1() == 0){
close(1);
dup(p[1]);   //part (a)
close(p[0]);
close(p[1]);
runcmd(pcmd->left);
}
if(fork1() == 0){
close(0);
dup(p[0]);
close(p[0]);
close(p[1]);
runcmd(pcmd->right);
}
close(p[0]);
close(p[1]);  //part (b)
wait();
wait();
```

Assume the shell gets the command "echo hello | grep hello". Let P denote the parent shell process that implements this command, and let CL and CR denote the child processes created to execute the left and right commands of the above pipe command respectively. Assume P has no other file descriptors open, except the standard input/output/error pointing to the console. Below are shown the various (global and per-process) open file tables right after P executes the `pipe` system call, but before it forks CL and CR.

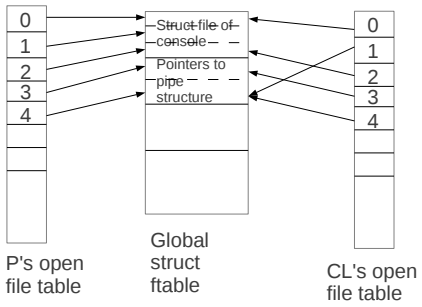


Draw similar figures showing the states of the various open file tables after the execution of lines marked (a) and (b) above. For each of parts (a) and (b), you must clearly draw both the global and per-process file tables, and illustrate the various pointers clearly. You may assume that all created child processes are still running by the time the line marked part (b) above is executed. You may

also assume that the scheduler switches to the child right after fork, so that the line marked part (a) in the child runs before the line marked part (b) in the parent.

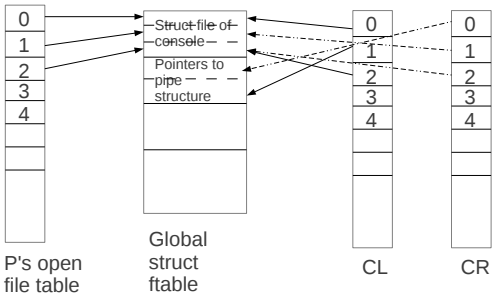
Ans:

(a) The fd 1 of the child process has been modified to point to one of the pipe's file descriptors



by the close and dup operations.

(b) CL and CR are now connected to either ends of the pipe. The parent has no pointers to the



pipe any more.

13. Consider a simple xv6-like filesystem, with the following deviations from the xv6 design. The filesystem uses a simple inode structure with only direct pointers (and no indirect blocks) to track file data blocks. Unlike xv6, this filesystem does not use logging or any other mechanism to guarantee crash consistency. The disk buffer cache follows the write-back design paradigm. Much like xv6, free data blocks and free inodes on disk are tracked via bitmaps. A process on this system performed a `write` system call on an open file, to add a new block of data at the end of the file. The successful execution of this system call changed several data and metadata blocks in the disk buffer cache. However, a power failure occurs before these changed blocks can be flushed to disk, causing changes to some disk blocks to be lost before being propagated to disk.

    (a) Consider the following scenario after the crash. The user reboots the system, and opens the file he wrote to just before the crash. From the file size on disk, it appears that his last write completed successfully. However, upon reading back the data he had written in the last block, he finds the data to be incorrect (garbage). Can you explain how this scenario can occur? Specifically, can you state which changed blocks pertaining to this system call were propagated to disk, and which weren't?

    Blocks correctly changed on disk:

    Blocks whose changes were lost during the crash:

    (b) Repeat the above question when the user finds himself in the following scenario. Immediately after reboot, the user finds that the data he had written to the file just before the crash did survive the crash, and he is able to read back the correct data from the file. However, after a few hours of using his system, he finds that the file now has incorrect data in the last block that he wrote before the crash.

    Blocks correctly changed on disk:

    Blocks whose changes were lost during the crash:

    (c) Now, suppose the filesystem implementation is changed in such a way that, for any system call, changes to data blocks are always written to the disk before changes to metadata blocks. Which of the two scenarios of parts (a) and (b) could still have occurred after this change to the filesystem, and which could have been prevented?

    Scenario(s) that still occur:

    Scenario(s) prevented:

    (d) Now, suppose the filesystem comes with a filesystem checker tool like `fsck`, which checks the consistency of filesystem metadata after a crash, and fixes any inconsistencies it finds. If this tool is run on the filesystem after the crash, which of the two scenarios of parts (a) and (b) can still happen, and which could have been prevented?

    Scenario(s) that still occur:

    Scenario(s) prevented:

    Ans: (a) inode and bitmap changed, data block change lost (b) inode and data block changed, bitmap change lost (causing block to be reused) (c) part a wont happen, b will still occur (d) part a can still happen, b wont occur

14. Which of the following system calls in xv6, when executed successfully, always result in a new entry being added to the open file table?

    (A) `open` (B) `dup` (C) `pipe` (D) `fork`

Ans: AC

15. Consider an in-memory inode pointer `struct inode *ip` in xv6 that is returned via a call to `iget`. Which of the following statements about this inode pointer is/are true?

    (A) The pointer returned from `iget` is an exclusive pointer, and another process that requests a pointer to the same inode will block until the previous process releases it via `iput`
    (B) The pointer returned by `iget` is non-exclusive, and multiple processes can obtain a pointer to the same inode via calls to `iget`
    (C) The contents of the inode pointer returned by `iget` are always guaranteed to be correct, and consistent with the information in the on-disk inode
    (D) The reference count of the inode returned by `iget` is always non-zero, i.e., `ip->ref > 0`

    Ans: BD

16. When is an inode marked as free on the disk in xv6?

    (A) As soon as its link count hits 0, even if the reference count of the in-memory inode is non-zero
    (B) As soon as the reference count of the in-memory inode hits 0, even if the link count of the inode is non-zero
    (C) When both the link count and reference count of the in-memory inode are 0
    (D) Cannot say; the answer depends on the exact sequence of system calls executed

    Ans: C

Lectures on Operating Systems (Mythili Vutukuru, IIT Bombay)

Practice Problems: Memory Management in xv6

1. Consider a system with V bytes of virtual address space available per process, running an xv6-like OS. Much like with xv6, low virtual addresses, up to virtual address U, hold user data. The kernel is mapped into the high virtual address space of every process, starting at address U and upto the maximum V. The system has P bytes of physical memory that must all be usable. The first K bytes of the physical memory holds the kernel code/data, and the rest $P - K$ bytes are free pages. The free pages are mapped once into the kernel address space, and once into the user part of the address space of the process they are assigned to. Like in xv6, the kernel maintains page table mappings for the free pages even after they have been assigned to user processes. The OS does not use demand paging, or any form of page sharing between user space processes. The system must be allowed to run up to N processes concurrently.

   (a) Assume $N = 1$. Assume that the values of V, U, and K are known for a system. What values of P (in terms of V, U, K) will ensure that all the physical memory is usable?

   (b) Assume the values of V, K, and N are known for a system, but the value of P is not known apriori. Suggest how you would pick a suitable value (or range of values) for U. That is, explain how the system designer must split the virtual address space into user and kernel parts.

   Ans:

   (a) The kernel part of the virtual address space of a process should be large enough to map all physical memory, so $V - U \geq P$. Further, the user part of the virtual address space of a process should fit within the free physical memory that is left after placing the kernel code, so $U \leq P - K$. Putting these two equations together will get you a range for P.

   (b) If there are N processes, the second equation above should be modified so that the combined user part of the N processes can fit into the free physical pages. So we will have $N * U \leq P - K$. We also have $P \leq V - U$ as before. Eliminating P (unknown), we get $U \leq \frac{V-K}{N+1}$.

2. Consider a system running the xv6 OS. A parent process P has forked a child C, after which C executes the `exec` system call to load a different binary onto its memory image. During the execution of `exec`, does the kernel stack of C get reinitialized or reallocated (much like the page tables of C)? If it does, explain what part of `exec` performs the reinitialization. If not, explain why not.

   Ans: The kernel stack cannot be reallocated during `exec`, because the kernel code is executing on the kernel stack itself, and releasing the stack on which the kernel is running would be disastrous. Small changes are made to the trap frame however, to point to the start of the new executable.

3. The xv6 operating system does not implement copy-on-write during fork. That is, the parent's user memory pages are all cloned for the child right at the beginning of the child's creation. If xv6 were to implement copy-on-write, briefly explain how you would implement it, and what changes need to be made to the xv6 kernel. Your answer should not just describe what copy-on-write is (do not say things like "copy memory only when parent or child modify it"), but instead concretely explain *how* you would ensure that a memory page is copied only when the parent/child wishes to modify it.

   Ans: The memory pages shared by parent and child would be marked read-only in the page table. Any attempt to write to the memory by the parent or child would trap the OS, at which point a copy of the page can be made.

4. Consider a process P in xv6 that has executed the `kill` system call to terminate a victim process V. If you read the implementation of `kill` in xv6, you will see that V is not terminated immediately, nor is its memory reclaimed during the execution of the `kill` system call itself.

   (a) Give one reason why V's memory is not reclaimed during the execution of `kill` by P.

   (b) Describe when V is actually terminated by the kernel.

   Ans: Memory cannot be reclaimed during the kill itself, because the victim process may actually be executing on another core. Processes are periodically checked for whether they have been killed (say, when they enter/exit kernel mode), and termination and memory reclamation happens at a time that is convenient to the kernel.

5. Consider the implementation of the `exec` system call in xv6. The implementation of the system call first allocates a new set of page tables to point to the new memory image, and switches page tables only towards the end of the system call. Explain why the implementation keeps the old page tables intact until the end of `exec`, and not rewrite the old page tables directly while building the new memory image.

   Ans: The `exec` system call retains the old page tables, so that it can switch back to the old image and print an error message if exec does not succeed. If exec succeeds however, the old memory image will no longer be needed, hence the old page tables are switched and freed.

6. In a system running xv6, for every memory access by the CPU, the function `walkpgdir` is invoked to translate the logical address to physical address. [T/F]

   Ans: F

7. Consider a process in xv6 that makes the `exec` system call. The EIP of the `exec` instruction is saved on the kernel stack of the process as part of handling the system call. When and under what conditions is this EIP restored from the stack causing the process to execute the statement after `exec`?

   Ans: If some error occurs during exec, the process uses the eip on trap frame to return to instruction after exec in the old memory image.

8. Consider a process in an xv6 system. Consider the following statement: "All virtual memory addresses starting from 0 to $N - 1$ bytes, where N is the process size (`proc->sz`), can be read by the process in user mode." Is the above statement true or false? If true, explain why. If false, provide a counter example.

   Ans: False, the guard page is not accessible by user.

9. When an xv6 process invokes the `exec` system call, where are the arguments to the system call first copied to, before the system call execution begins? Tick one: user heap / user stack / trap frame / context structure

   Ans: user stack

10. When a process successfully returns from the `exec` system call to its new memory image in xv6, where are the commandline arguments given to the new executable to be found? Tick one: user heap / user stack / trap frame / context structure

    Ans: user stack

11. After the `exec` system call completes successfully in xv6, where is the EIP of the starting instruction of the new executable stored, to enable the process to start execution at the entry of the new code? Tick one: user heap / user stack / trap frame / context structure

    Ans: trap frame

12. Consider the list of free memory frames maintained by xv6 in the free list. Whenever a process in kernel mode requests memory via the function kalloc(), xv6 extracts and returns free frames from this list. Which of the following things can be stored in the pages thus allocated from the free list?

    (a) New page table created for child process during fork

    (b) New memory image (code/data/stack/heap) created for child during fork

    (c) Kernel stack of new process created in fork

    (d) New page table created for the new memory image in exec

    Ans: (a), (b), (c), (d)

13. Consider a process P in xv6 that executes the sbrk system call to increase the size of the user part of its virtual address space from N1 bytes to N2 bytes. Assume N1 and N2 are multiples of page size, N2 ¿ N1, and the difference between N1 and N2 is K pages. Which of the following statements is/are true about the actions that occur during the execution of the system call?

    (a) The OS assigns K free physical frames to the process and adds the frame numbers into the page table.

    (b) The OS does not assign any new physical frames to the process, but updates the page table.

    (c) The OS updates (N2-N1) page table entries in the page table of P.

    (d) The OS updates K page table entries in the page table of P.

    Ans: (a), (d)

14. Consider a process P running in xv6. The high virtual addresses in the address space of P are assigned to OS code/data. Consider a virtual address V assigned to OS code/data. Which of the following statements is/are true?

    (a) If the CPU accesses address V in user mode, the MMU raises a trap to the OS.

    (b) The address V is translated to the same physical address by the page tables of all processes in the system.

(c) The address V can be translated to different physical address by the page tables of different processes in the system.

(d) The page table entry that translates address V to a physical address is present in the page table of P only when it is running in kernel mode.

Ans: (a), (b)

15. Consider a process running in xv6. The page table of the process maps virtual address V to physical address P. Which of the following statements is/are true? Assume KERNBASE is set to 2GB in xv6.

(a) V = P + KERNBASE always

(b) V = P – KERNBASE always

(c) V = P + KERNBASE only for V ¿= KERNBASE

(d) V = P + KERNBASE only for V ¡ KERNBASE

Ans: (c)

Lectures on Operating Systems (Mythili Vutukuru, IIT Bombay)

Practice Problems: Process Management in xv6

1. Consider the following lines of code in a program running on xv6.

```
int ret = fork();
if(ret==0) { //do something in child}
else { //do something in parent}
```

   (a) When a new child process is created as part of handling fork, what does the kernel stack of the new child process contain, after fork finishes creating it, but just before the CPU switches away from the parent?

   (b) How is the kernel stack of the newly created child process different from that of the parent?

   (c) The EIP value that is present in the trap frames of the parent and child processes decides where both the processes resume execution in user mode. Do both the EIP pointers in the parent and child contain the same logical address? Do they point to the same physical address in memory (after address translation by page tables)? Explain.

   (d) How would your answer to (c) above change if xv6 implemented copy-on-write during fork?

   (e) When the child process is scheduled for the first time, where does it start execution in kernel mode? List the steps until it finally gets to executing the instruction after fork in the program above in user mode.

   Ans:

   (a) It contains a trap frame, followed by the context structure.

   (b) The parent's kernel stack has only a trap frame, since it is still running and has not been context switched out. Further, the value of the EAX register in the two trap frames is different, to return different values in parent and child.

   (c) The EIP value points to the same logical address, but to different physical addresses, as the parent and child have different memory images.

   (d) With copy-on-write, the physical addresses may be the same as well, as long as both parent and child have not modified anything.

   (e) Starts at forkret, followed by trapret. Pops the trapframe and starts executing at the instruction right after fork.

2. Suppose a machine (architecture: x86, single core) has two runnable processes P1 and P2. P1 executes a line of code to read 1KB of data from an open file on disk to a buffer in its memory.

The content requested is not available in the disk buffer cache and must be fetched from disk. Describe what happens from the time the instruction to read data is started in P1, to the time it completes (causing the process to move on to the next instruction in the program), by answering the following questions.

(a) The code to read data from disk will result in a system call, and will cause the x86 CPU to execute the `int` instruction. Briefly describe what the CPU's `int` instruction does.

(b) The `int` instruction will then call the kernel's code to handle the system call. Briefly describe the actions executed by the OS interrupt/trap/system call handling code before the read system call causes P1 to block.

(c) Now, because process P1 has made a blocking system call, the CPU scheduler context switches to some other process, say P2. Now, the data from the disk that unblocks P1 is ready, and the disk controller raises an interrupt while P2 is running. On whose kernel stack does this interrupt processing run?

(d) Describe the contents of the kernel stacks of P1 and P2 when this interrupt is being processed.

(e) Describe the actions performed by P2 in kernel mode when servicing this disk interrupt.

(f) Right after the disk interrupt handler has successfully serviced the interrupt above, and before any further calls to the scheduler to context switch from P2, what is the state of process P1?

Ans:

(a) The CPU switches to kernel mode, switches to the kernel stack of the process, and pushes some registers like the EIP onto the kernel stack.

(b) The kernel pushes a few other registers, updates segment registers, and starts executing the system call code, which eventually causes P1 to block.

(c) P2's kernel stack

(d) P2's kernel stack has a trapframe (since it switched to kernel mode). P1's kernel stack has both a context structure and a trap frame (since it is currently context switched out).

(e) P2 saves user context, switches to kernel mode, services the disk interrupt that unblocks P1, marks P1 as ready, and resumes its execution in userspace.

(f) Ready / runnable.

3. Consider a newly created process in xv6. Below are the several points in the code that the new process passes through before it ends up executing the line just after the fork statement in its user space. The EIP of each of these code locations exists on the kernel stack when the process is scheduled for the first time. For each of these locations below, precisely explain where on the kernel stack the corresponding EIP is stored (in which data structure etc.), and when (as part of which function or stage of process creation) is the EIP written there. Be as specific as possible in your answers.

(a) `forkret`

(b) `trapret`

(c) just after the `fork()` system call in userspace

Ans:

(a) EIP of forkret is stored in struct context by allocproc.

(b) EIP of trapret is stored on kernel stack by allocproc.

(c) EIP of fork system call code is stored in trapframe in parent, and copied to child's kernel stack in the fork function.

4. Consider a process that has performed a blocking disk read, and has been context switched out in xv6. Now, when the disk interrupt occurs with the data requested by the process, the process is unblocked and context switched in immediately, as part of handling the interrupt. [T/F]

   Ans: F

5. In xv6, state the system call(s) that result in new `struct proc` objects being allocated.

   Ans: fork

6. Give an example of a scenario in which the xv6 dispatcher / `swtch` function does NOT use a `struct context` created by itself previously, but instead uses an artificially hand-crafted `struct context`, for the purpose of restoring context during a context switch.

   Ans: When running process for first time (say, after fork).

7. Give an example of a scenario in xv6 where a `struct context` is stored on the kernel stack of a process, but this context is never used or restored at any point in the future.

   Ans: When process has finished after exit, its saved context is never restored.

8. Consider a parent process P that has executed a fork system call to spawn a child process C. Suppose that P has just finished executing the system call code, but has not yet returned to user mode. Also assume that the scheduler is still executing P and has not context switched it out. Below are listed several pieces of state pertaining to a process in xv6. For each item below, answer if the state is identical in both processes P and C. Answer Yes (if identical) or No (if different) for each question.

   (a) Contents of the PCB (struct proc). That is, are the PCBs of P and C identical? (Yes/No)

   (b) Contents of the memory image (code, data, heap, user stack etc.).

   (c) Contents of the page table stored in the PCB.

   (d) Contents of the kernel stack.

   (e) EIP value in the trap frame.

   (f) EAX register value in the trap frame.

   (g) The physical memory address corresponding to the EIP in the trap frame.

   (h) The files pointed at by the file descriptor table. That is, are the file structures pointed at by any given file descriptor identical in both P and C?

   Ans:

   (a) No

   (b) Yes

   (c) No

(d) No

(e) Yes

(f) No

(g) No

(h) Yes

9. Suppose the kernel has just created the first user space "init" process, but has not yet scheduled it. Answer the following questions.

   (a) What does the EIP in the trap frame on the kernel stack of the process point to?

   (b) What does the EIP in the context structure on the kernel stack (that is popped when the process is context switched in) point to?

   Ans:

   (a) address 0 (first line of code in init user code)

   (b) forkret / trapret

10. Consider a process P that forks a child process C in xv6. Compare the trap frames on the kernel stacks of P and C just after the fork system call completes execution, and before P returns back to user mode. State one difference between the two trap frames at this instant. Be specific in your answer and state the exact field/register value that is different.

    Ans: EAX register has different value.

11. In xv6, the EIP within the `struct context` on the kernel stack of a process usually points to the `swtch` statement in the `sched` function, where the process gives up its CPU and switches to the scheduler thread during a context switch. Which processes are an exception to this statement? That is, for which processes does the EIP on the context structure point to some other piece of code?

    Ans: Newly created processes / processes running for first time

12. When an trap occurs in xv6, and a process shifts from user mode to kernel mode, which entity switches the CPU stack pointer from pointing to the user stack of the running program to its kernel stack? Tick one: x86 hardware instruction / xv6 assembly code

    Ans: x86 hardware

13. Consider a process P in xv6, which makes a system call, goes to kernel mode, runs the system call code, and comes back into user mode again. The value of the EAX register is preserved across this transition. That is, the value of the EAX register just before the process started the system call will always be equal to its value just after the process has returned back to user mode. [T/F]

    Ans: False, EAX is used to store system call number and return value, so it changes.

14. When a trap causes a process to shift from user mode to kernel mode in xv6, which CPU data registers are stored in the trapframe (on the kernel stack) of the process? Tick one: all registers / only callee-save registers

    Ans: All registers

15. When a process in xv6 wishes to pass one or more arguments to the system call, where are these arguments initially stored, before the process initiates a jump into kernel mode? Tick one: user stack / kernel stack

    Ans: User stack, as user program cannot access kernel stack

16. Consider the context switch of a CPU from the context of process P1 to that of process P2 in xv6. Consider the following two events in the chronological order of the events during the context switch: (E1) the ESP (stack pointer) shifts from pointing to the kernel stack of P1 to the kernel stack of P2; (E2) the EIP (program counter) shifts from pointing to an address in the memory allocated to P1 to an address in the memory allocated to P2. Which of the following statements is/are true regarding the relative ordering of events E1 and E2?

    (a) E1 occurs before E2.

    (b) E2 occurs before E1.

    (c) E1 and E2 occur simultaneously via an atomic hardware instruction.

    (d) The relative ordering of E1 and E2 can vary from one context switch to the other.

    Ans: (a)

17. Consider the following actions that happen during a context switch from thread/process P1 to thread/process P2 in xv6. (One of P1 or P2 could be the scheduler thread as well.) Arrange the actions below in chronological order, from earliest to latest.

    (A) Switch ESP from kernel stack of P1 to that of P2

    (B) Pop the callee-save registers from the kernel stack of P2

    (C) Push the callee-save registers onto the kernel stack of P1

    (D) Push the EIP where execution of P1 stops onto the kernel stack of P1.

    Ans: DCAB

18. Consider a process P in xv6 that invokes the wait system call. Which of the following statements is/are true?

    (a) If P does not have any zombie children, then the wait system call returns immediately.

    (b) The wait system call always blocks process P and leads to a context switch.

    (c) If P has exactly one child process, and that child has not yet terminated, then the wait system call will cause process P to block.

    (d) If P has two or more zombie children, then the wait system call reaps all the zombie children of P and returns immediately.

    Ans: (c)

19. Consider a process P in xv6 that executes the exec system call successfully. Which of the following statements is/are true?

    (a) The exec system call changes the PID of process P.

    (b) The exec system call allocates a new page table for process P.

(c) The exec system call allocates a new kernel stack for process P.

(d) The exec system call changes one or more fields in the trap frame on the kernel stack of process P.

Ans: (b), (d)

20. Consider a process P in xv6 that executes the exec system call successfully. Which of the following statements is/are true?

    (a) The arguments to the exec system call are first placed on the user stack by the user code.

    (b) The arguments to the exec system call are first placed on the kernel stack by the user code.

    (c) The arguments (argc, argv) to the new executable are placed on the kernel stack by the exec system call code.

    (d) The arguments (argc, argv) to the new executable are placed on the user stack by the exec system call code.

    Ans: (a), (d)

21. Consider a newly created child process C in xv6 that is scheduled for the first time. At the point when the scheduler is just about to context switch into C, which of the following statements is/are true about the kernel stack of process C?

    (a) The top of the kernel stack contains the context structure, whose EIP points to the instruction right after the fork system call in user code.

    (b) The bottom of the kernel stack has the trapframe, whose EIP points to the forkret function in OS code.

    (c) The top of the kernel stack contains the context structure, whose EIP points to the forkret function in OS code.

    (d) The bottom of the kernel stack contains the trap frame, whose EIP points to the trapret function in OS code.

    Ans: (c)

22. Consider a trapframe stored on the kernel stack of a process P in xv6 that jumped from user mode to kernel mode due to a trap. Which of the following statements is/are true?

    (a) All fields of the trapframe are pushed onto the kernel stack by the OS code.

    (b) All fields of the trapframe are pushed onto the kernel stack by the x86 hardware.

    (c) The ESP value stored in the trapframe points to the top of the kernel stack of the process.

    (d) The ESP value stored in the trapframe points to the top of the user stack of the process.

    Ans: (d)

23. Consider a process P that has made a blocking disk read in xv6. The OS has issued a disk read command to the disk hardware, and has context switched away from P. Which of the following statements is/are true?

(a) The top of the kernel stack of P contains the return address, which is the value of EIP pointing to the user code after the read system call.

(b) The bottom of the kernel stack of P contains the trapframe, whose EIP points to the user code after the read system call.

(c) The top of the kernel stack of P contains the context structure, whose EIP points to the user code after the read system call.

(d) The CPU scheduler does not run P again until after the disk interrupt that unblocks P is raised by the device hardware.

Ans: (b), (d)

24. In the implementation of which of the following system calls in xv6 are new ptable entries allocated or old ptable entries released (marked as unused)?

(a) fork

(b) exit

(c) exec

(d) wait

Ans: (a), (d)

25. A process has invoked exit() in xv6. The CPU has completed executing the OS code corresponding to the exit system call, and is just about to invoke the swtch() function to switch from the terminated process to the scheduler thread. Which of the following statements is/are true?

(a) The stack pointer ESP is pointing to some location within the kernel stack of the terminated process

(b) The MMU is using the page table of the terminated process

(c) The state of the terminated process in the ptable is RUNNING

(d) The state of the terminated process in the ptable is ZOMBIE

Ans: (a), (b), (d)

Lectures on Operating Systems (Mythili Vutukuru, IIT Bombay)

Practice Problems: Synchronization in xv6

1. Modern operating systems disable interrupts on specific cores when they need to turn off preemption, e.g., when holding a spin lock. For example, in xv6, interrupts can be disabled by a function call `cli()`, and reenabled with a function call `sti()`. However, functions that need to disable and enable interrupts do not directly call the `cli()` and `sti()` functions. Instead, the xv6 kernel disables interrupts (e.g., while acquiring a spin lock) by calling the function `pushcli()`. This function calls `cli()`, but also maintains of count of how many push calls have been made so far. Code that wishes to enable interrupts (e.g., when releasing a spin lock) calls `popcli()`. This function decrements the above push count, and enables interrupts using `sti()` only after the count has reached zero. That is, it would take two calls to `popcli()` to undo the effect of two `pushcli()` calls and restore interrupts. Provide one reason why modern operating systems use this method to disable/enable interrupts, instead of directly calling the `cli()` and `sti()` functions. In other words, explain what would go wrong if every call to `pushcli()` and `popcli()` in xv6 were to be replaced by calls to `cli()` and `sti()` respectively.

   Ans: If one acquires multiple spinlocks (say, while serving nested interrupts, or for some other reason), interrupts should be enabled only after locks have been released. Therefore, the push and pop operations capture how many times interrupts have been disabled, so that interrupts can be reenabled only after all such operations have been completed.

2. Consider an operating system where the list of process control blocks is stored as a linked list sorted by pid. The implementation of the wakeup function (to wake up a process waiting on a condition) looks over the list of processes in order (starting from the lowest pid), and wakes up the first process that it finds to be waiting on the condition. Does this method of waking up a sleeping process guarantee bounded wait time for every sleeping process? If yes, explain why. If not, describe how you would modify the implementation of the wakeup function to guarantee bounded wait.

   Ans: No, this design can have starvation. To fix it, keep a pointer to where the wakeup function stopped last time, and continue from there on the next call to wakeup.

3. Consider an operating system that does not provide the `wait` system call for parent processes to reap dead children. In such an operating system, describe one possible way in which the memory allocated to a terminated process can be reclaimed correctly. That is, identify one possible place in the kernel where you would put the code to reclaim the memory.

   Ans: One possible place is the scheduler code itself: while going over the list of processes, it can identify and clean up zombies. Note that the cleanup cannot happen in the exit code itself, as the process memory must be around till it invokes the scheduler.

4. Consider a process that invokes the `sleep` function in xv6. The process calling sleep provides a lock `lk` as an argument, which is the lock used by the process to protect the atomicity of its call to sleep. Any process that wishes to call `wakeup` will also acquire this lock `lk`, thus avoiding a call to wakeup executing concurrently with the call to sleep. Assume that this lock `lk` is not `ptable.lock`. Now, if you recall the implementation of the sleep function, the lock `lk` is released before the process invokes the scheduler to relinquish the CPU. Given this fact, explain what prevents another process from running the wakeup function, while the first process is still executing sleep, after it has given up the lock `lk` but before its call to the scheduler, thus breaking the atomicity of the sleep operation. In other words, explain why this design of xv6 that releases `lk` before giving up the CPU is still correct.

   Ans: Sleep continues to hold ptable.lock even after releasing the lock it was given. And wakeup requires ptable.lock. Therefore, wakeup cannot execute concurrently with sleep.

5. Consider the `yield` function in xv6, that is called by the process that wishes to give up the CPU after a timer interrupt. The yield function first locks the global lock protecting the process table (`ptable.lock`), before marking itself as RUNNABLE and invoking the scheduler. Describe what would go wrong if `yield` locked `ptable.lock` AFTER setting its state to RUNNABLE, but before giving up the CPU.

   Ans: If marked runnable, another CPU could find this process runnable and start executing it. One process cannot run on two cores in parallel.

6. Provide one reason why a newly created process in xv6, running for the first time, starts its execution in the function `forkret`, and not in the function `trapret`, given that the function `forkret` almost immediately returns to `trapret`. In other words, explain the most important thing a newly created process must do before it pops the trap frame and executes the return from the trap in `trapret`.

   Ans: It releases ptable.lock and preserves the atomicity of the context switch.

7. Consider a process P in xv6 that acquires a spinlock L, and then calls the function `sleep`, providing the lock L as an argument to `sleep`. Under which condition(s) will lock L be released *before* P gives up the CPU and blocks?

   (a) Only if L is `ptable.lock`

   (b) Only if L is not `ptable.lock`

   (c) Never

   (d) Always

   Ans: (b)

8. Consider a system running xv6. You are told that a process in kernel mode acquires a spinlock (it can be any of the locks in the kernel code, you are not told which one). While the process holds this spin lock, is it correct OS design for it to:

   (a) process interrupts on the core in which it is running?

   (b) call the `sched` function to give up the CPU?

For each question above, you must first answer Yes or No. If your answer is yes, you must give an example from the code (specify the name of the lock, and any other information about the scenario) where such an event occurs. If your answer is no, explain why such an event cannot occur.

Ans:

(a) No, it cannot. The interrupt may also require the same spinlock, leading to a deadlock.

(b) Yes it is possible. Processes giving up CPU call `sched` with `ptable.lock` held.

9. Consider the following snippet of code from the `sleep` function of xv6. Here, `lk` is the lock given to the sleep function as an argument.

```
if(lk != &ptable.lock){
  acquire(&ptable.lock);
  release(lk);
}
```

For each of the snippets of code shown below, explain what would happen if the original code shown above were to be replaced by the code below. Does this break the functionality of `sleep`? If yes, explain what would go wrong. If not, explain why not.

(a)
```
acquire(&ptable.lock);
release(lk);
```

(b)
```
release(lk);
acquire(&ptable.lock);
```

Ans:

(a) This code will deadlock if the lock given to sleep is `ptable.lock` itself.

(b) A wakeup may run between the release and acquire steps, leading to a missed wakeup.

10. In xv6, when a process calls `sleep` to block on a disk read, suggest what could be used as a suitable channel argument to the `sleep` function (and subsequently by `wakeup`), in order for the sleep and wakeup to happen correctly.

Ans: Address of struct buf (can be block number also?)

11. In xv6, when a process calls `wait` to block for a dead child, suggest what could be used as a suitable channel argument in the `sleep` function (and subsequently by `wakeup`), in order for the sleep and wakeup to happen correctly.

Ans: Address of parent struct proc (can be PID of parent?)

12. Consider the exit system call in xv6. The exit function acquires ptable.lock before giving up the CPU (in the function sched) for one last time. Who releases this lock subsequently?

Ans: The process that runs immediately afterwards (or scheduler)

13. In xv6, when a process calls wakeup on a channel to wakeup another process, does this lead to an immediate context switch of the process that called wakeup (immediately after the wakeup instruction)? (Yes/No)

    Ans: No

14. When a process terminates in xv6, when is the struct proc entry of the process marked as unused/free?

    (a) During the execution of exit

    (b) During the sched function that performs the context switch

    (c) In the scheduler, when it iterates over the array of all struct proc

    (d) During the execution of wait by the parent

    Ans: (d)

15. In which of the following xv6 system call implementations is there a possibility of the process calling `sched()` to give up the CPU? Assume no timer interrupts occur during the system call execution. Tick all that apply: `fork` / `exit` / `wait` / none of the above

    Ans: exit, wait

16. In which of the following xv6 system call implementations will a process *always* invoke `sched()` to give up the CPU? Assume no timer interrupts occur during the system call execution. Tick all that apply: `fork` / `exit` / `wait` / none of the above

    Ans: exit

17. Under which conditions does the `wait()` system call in xv6 block? Tick all that apply: when the process has [no children / only one running child / only one zombie child / two children, one running and one a zombie]

    Ans: Only one running child

18. Consider a system with two CPU cores (C0 and C1) that is running the xv6 OS. A process P0 running on core C0 is in kernel mode, and has acquired a kernel spinlock. Another process P1 on core C1 is also in kernel mode, and is busily spinning for the same spinlock that is held by P0. Which of the following best describes the set of cores on which interrupts are disabled?

    (A) C0 and C1 (B) Only C0 (C) Only C1 (D) Neither C0 nor C1

    Ans: A, interrupts need to be disabled before starting to spin also.

19. Consider a process in xv6 that invokes the wakeup function in kernel mode. Which of the following statements is/are true?

    (a) The wakeup function immediately causes a context switch to one of the processes sleeping on a particular channel value.

    (b) The wakeup function wakes up (marks as ready) all processes sleeping on a particular channel value.

    (c) The wakeup function wakes up (marks as ready) only the first process that is blocked on a particular channel value.

(d) The wakeup function wakes up (marks as ready) only the last process that is blocked on a particular channel value.

Ans: (b)

20. Consider a process in kernel mode in xv6, which invokes the sleep function to block. One of the arguments to the sleep function is a kernel spinlock. Which of the following statements is/are true?

    (a) If the lock given to sleep is ptable.lock, then the lock is not released before the context switch.

    (b) If the lock given to sleep is not ptable.lock, then the lock is not released before the context switch.

    (c) If the lock given to sleep is not ptable.lock, then the lock is released before the context switch, but only after acquiring ptable.lock.

    (d) If the lock given to sleep is not ptable.lock, then the lock is released before the context switch and ptable.lock need not be acquired.

    Ans: (a), (c)