Semaphores (Dijkstra, 1962)

Programming construct for synchronization:
- New data type: semaphore, an integer \( s \)
  
  \[
  \text{semaphore } s = K; \quad \text{// initialize } s \text{ to } K
  \]
- New operations (defined on semaphores):
  - \text{wait}(\text{semaphore } s)
  - \text{signal}(\text{semaphore } s)
  \[
  s = s + 1 \quad \text{(one waiting thread may now be able to proceed)}
  \]
- Semantic guarantee: A semaphore initialized to \( K \) enforces the precedence constraint:

\[
\text{signal}(s) \prec \text{wait}(s) + K
\]

The \( i \)-th call to \text{signal}(s) must complete before the \( (i+K)\)-th call to \text{wait}(s) completes.

Semaphores for Precedence

\[
\text{semaphore } s = 0;
\]

Goal: Want statement \( A_2 \) in thread A to complete before statement \( B_4 \) in thread B begins.

<table>
<thead>
<tr>
<th>Thread A</th>
<th>Thread B</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_1 )</td>
<td>( B_1 )</td>
</tr>
<tr>
<td>( A_2 )</td>
<td>( B_2 )</td>
</tr>
<tr>
<td>( A_3 )</td>
<td>( B_3 )</td>
</tr>
<tr>
<td>( A_4 )</td>
<td>( B_4 )</td>
</tr>
<tr>
<td>( A_5 )</td>
<td>( B_5 )</td>
</tr>
</tbody>
</table>

Recipe:
- Declare \( s = 0 \)
- \text{signal}(s) at start of arrow
- \text{wait}(s) at end of arrow

Semaphores for Resource Allocation

Abstract problem:
- Pool of K resources
- Many threads, each needs resource for occasional uninterrupted period
- Must guarantee that at most K resources are in use at any time

Solution using semaphores:
- In shared memory:
  \[
  \text{semaphore } s = K; \quad \text{// K resources}
  \]
- Using resources:
  \[
  \text{wait}(s); \quad \text{// Allocate a resource}
  \]
  \[
  \text{use it for a while}
  \]
  \[
  \text{signal}(s); \quad \text{// return it to pool}
  \]

Invariant: Semaphore value = number of resources left in pool

Semaphores for Mutual Exclusion

\[
\text{semaphore lock} = 1;
\]

void \text{debit(int account, int amount)}

\[
\text{wait(lock); // Wait for exclusive access}
\]

\[
\text{balance[account]} = t - \text{amount;}
\]

\[
\text{signal(lock); // Finished with lock}
\]

Lock controls access to critical section
Issue: Lock granularity
- One lock for all accounts?
- One lock per account?
- One lock for all accounts ending in 004?

Synchronization: The Dark Side

The naïve use of synchronization constructs can introduce its own set of problems, particularly when a thread requires access to more than one protected resource.

void \text{transfer(int account1, int account2, int amount)}

\[
\text{wait(lock[account1]);}
\]

\[
\text{balance[account1]} = \text{balance[account1]} - \text{amount;}
\]

\[
\text{wait(lock[account2]);}
\]

\[
\text{balance[account2]} = \text{balance[account2]} + \text{amount;}
\]

\[
\text{signal(lock[account1]);}
\]

\[
\text{signal(lock[account2]);}
\]

What can go wrong here?
- Thread 1: \( \text{wait(lock[account1])} \);
- Thread 2: \( \text{wait(lock[account2])} \);
- Cannot complete

Example: Dealing With Deadlocks

Can you fix the transfer method to avoid deadlock?

void \text{transfer(int account1, int account2, int amount)}

\[
\text{int } a = \text{min(account1, account2);}
\]

\[
\text{int } b = \text{max(account1, account2);}
\]

\[
\text{wait(lock[a]);}
\]

\[
\text{balance[account1]} = \text{balance[account1]} - \text{amount;}
\]

\[
\text{wait(lock[b]);}
\]

\[
\text{balance[account2]} = \text{balance[account2]} + \text{amount;}
\]

\[
\text{signal(lock[a]);}
\]

\[
\text{signal(lock[b]);}
\]

No thread can make progress → Deadlock
Problem 1.

Schro Dinger has a company that produces pairs of entangled particles, which are then packaged and sent to manufacturers of quantum computers. Since it’s a complicated process, there are multiple machines that produce particle pairs; each machine runs the Producer code shown below.

The completed particle pairs are placed in the particle buffer, where they take up 2 of the buffer locations. There’s a single packaging machine that takes a particle pair from the particle buffer and prepares it for shipment; the packing machine runs the Consumer code shown below.

To prevent any violations of the boundary conditions the following rules must be followed:

1. A production machine can only place a particle pair in the buffer if there are two spaces available.
2. The particle pair must be stored in consecutive buffer locations, i.e., a particle from some other production machine can’t appear between the particles that make up the pair.
3. The capacity of the buffer (100 particles, or 50 particle pairs) can’t be exceeded.
4. The packaging machine breaks if it accesses the buffer and finds it empty – it should only proceed when there are at least two particles in the buffer.

Schro has heard of semaphores but is unsure how to use them to ensure the rules are followed.

- Please insert the appropriate semaphores, WAITs, and SIGNALs into the Producer and Consumer code to ensure correct operation and to prevent deadlock.
- Be sure to indicate initial values for any semaphores you use.
- Remember: there are multiple producers and a single consumer!
- For full credit, use a minimum number of semaphores and don’t introduce unnecessary precedence constraints.

```
Shared Memory
particle buffer[100];  // holds 100 particles
Semaphores and initial values: P=0, S=50, M=1
```

```
Producer
PLoop:
Produce pair P1, P2
WAIT(S); WAIT(M)
Place P1 in buffer
Place P2 in buffer
SIGNAL(M); SIGNAL(P)
Go to PLoop

Consumer
CLoop:
WAIT(P)
Fetch P1 from buffer
FETCH P2 from buffer
SIGNAL(S)
Package and ship
Go to CLoop
```

P = # of particle pairs in buffer, enforces rule 4
S = # of pair spaces in buffer, enforces rules 1 and 3
M = mutual exclusion lock, enforces rule 2 (when there are multiple producers)
**Problem 2.**

The following three processes are run on a shared processor. They can coordinate their execution via shared semaphores that respond to the standard signal(S) and wait(S) procedures. Their intent is to print the word HELLO. Assume that execution may switch between any of the three processes at any point in time.

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
</tr>
</thead>
</table>
| Loop1: print(“H”)  
print(“E”)  
goto Loop1 | Loop2: print(“L”)  
goto Loop2 | Loop3: print(“O”)  
goto Loop3 |

(A) Assuming that no semaphores are being used, for each of the following sequences of characters, specify whether or not this system could produce that output.

- LEHO (YES/NO): ______
- HLOE (YES/NO): ______
- LOL (YES/NO): ______

*Need H before E*

(B) You would like to ensure that only the sequence HELLO can be printed and that it will be printed exactly once. Add any missing wait(S) and signal(S) calls to the code below (where S is one of a, b or c) to ensure that the three processes can only print HELLO exactly once. Remember to specify the **initial value** for each of your semaphores. *Recall that semaphores cannot be initialized to negative numbers.*
Semaphores: \( a = \_ \); \( b = \_ \); \( c = \_ \);

<table>
<thead>
<tr>
<th>Process 1</th>
<th>Process 2</th>
<th>Process 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop1:</td>
<td>Loop2:</td>
<td>Loop3:</td>
</tr>
<tr>
<td>( \text{wait}(a) )</td>
<td>( \text{wait}(b) )</td>
<td>( \text{wait}(c) )</td>
</tr>
<tr>
<td>( \text{print}(“H”) )</td>
<td>( \text{print}(“L”) )</td>
<td>( \text{print}(“O”) )</td>
</tr>
<tr>
<td>( \text{print}(“E”) )</td>
<td>( \text{signal(c)} )</td>
<td></td>
</tr>
<tr>
<td>( \text{signal(b)} )</td>
<td></td>
<td>( \text{signal(b)} )</td>
</tr>
<tr>
<td>goto Loop1</td>
<td>goto Loop2</td>
<td>goto Loop3</td>
</tr>
</tbody>
</table>
Problem 3.

The following pair of processes share the variable counter, which has been given an initial value of 10 before execution of either process begins and is stored at address 0x100:

<table>
<thead>
<tr>
<th>Process A</th>
<th>Process B</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>A1: lw x1, 0x100(x0)</td>
<td>B1: lw x1, 0x100(x0)</td>
</tr>
<tr>
<td>addi x1, 1, x1</td>
<td>addi x1, 2, x1</td>
</tr>
<tr>
<td>A2: sw x1, 0x100(x0)</td>
<td>B2: sw x1, 0x100(x0)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

(A) If Processes A and B are run on a timesharing system, there are six possible orders in which the LD and ST instructions might be executed. For each of the orderings, please give the final value of the counter variable.

- A1 A2 B1 B2: counter = __________
- B1 A1 B2 A2: counter = __________
- A1 B1 A2 B2: counter = __________
- B1 A1 A2 B2: counter = __________
- A1 B1 B2 A2: counter = __________
- B1 B2 A1 A2: counter = __________

In the following two questions you are asked to modify the original programs for processes A and B by adding the minimum number of semaphores and signal and wait operations to guarantee that the final result of executing the two processes will be a specific value for counter. Give the initial values for every semaphore you introduce. For full credit, your solution should allow all execution orders that result in the required value.

(B) Add semaphores (with initial values) so that the final value of counter is 12.

<table>
<thead>
<tr>
<th>Semaphores: X=0, Y=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process A</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>A1: lw x1, 0x100(x0)</td>
</tr>
<tr>
<td>addi x1, 1, x1</td>
</tr>
<tr>
<td>wait(x)</td>
</tr>
<tr>
<td>A2: sw x1, 0x100(x0)</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

(C) Add semaphores (with initial values), so that the final value of counter is not 13.

<table>
<thead>
<tr>
<th>Semaphores: X=0, Y=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process A</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>A1: lw x1, 0x100(x0)</td>
</tr>
<tr>
<td>signal(x)</td>
</tr>
<tr>
<td>addi x1, 1, x1</td>
</tr>
<tr>
<td>wait(y)</td>
</tr>
<tr>
<td>A2: sw x1, 0x100(x0)</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
Problem 4.

P1 and P2 are processes that run concurrently. P1 has two sections of code where section A is followed by section B. Similarly, P2 has two sections: C followed by D. Within each process execution proceeds sequentially, so we are guaranteed that A ≤ B, i.e., A precedes B. Similarly, we know that C ≤ D. There is no looping; each process runs exactly once. You will be asked to add semaphores to the programs – you may need to use more than one semaphore. Please give the initial values of any semaphores you use. For full credit use a minimum number of semaphores and don’t introduce any unnecessary precedence constraints.

(A) Please add WAIT(…) and SIGNAL(…) statements as needed in the spaces below so that the precedence constraint B ≤ C is satisfied, i.e., execution of P1 finishes before execution of P2 begins.

Add WAIT and SIGNAL statements so that B ≤ C

Semaphore initial values: S=0

<table>
<thead>
<tr>
<th>Process P1</th>
<th>Process P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>…Section A code…</td>
<td>wait(S)</td>
</tr>
<tr>
<td>…Section B code…</td>
<td>… Section C code…</td>
</tr>
<tr>
<td>signal(S)</td>
<td>…Section D code…</td>
</tr>
</tbody>
</table>
(B) Please add `WAIT(…)` and `SIGNAL(…)` statements as needed in the spaces below so that $D \leq A$ or $B \leq C$, i.e., executions of P1 and P2 cannot overlap, but are allowed to occur in either order.

**Add WAIT and SIGNAL statements so that $D \leq A$ or $B \leq C$**

Semaphore initial values: _______________

<table>
<thead>
<tr>
<th>Process P1</th>
<th>Process P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>wait(M)</td>
<td>wait(M)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>...Section A code...</td>
<td>... Section C code...</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>...Section B code...</td>
<td>...Section D code...</td>
</tr>
<tr>
<td>signal(M)</td>
<td>signal(M)</td>
</tr>
</tbody>
</table>

(C) Please add `WAIT(…)` and `SIGNAL(…)` statements as needed in the spaces below so that $A \leq D$ and $C \leq B$, i.e., the first section (A and C) of both processes completes execution before the second section (B or D) of either process begins execution.

**Add WAIT and SIGNAL statements so that $A \leq D$ and $C \leq B$**

Semaphore initial values: $S=\emptyset$, $T=\emptyset$

<table>
<thead>
<tr>
<th>Process P1</th>
<th>Process P2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>signal(S)</td>
</tr>
<tr>
<td></td>
<td>wait(T)</td>
</tr>
<tr>
<td>...Section A code...</td>
<td>... Section C code...</td>
</tr>
<tr>
<td>signal(S)</td>
<td>signal(T)</td>
</tr>
<tr>
<td>wait(T)</td>
<td>wait(S)</td>
</tr>
<tr>
<td>...Section B code...</td>
<td>...Section D code...</td>
</tr>
</tbody>
</table>
Problem 5.

The MIT Safety Office is worried about congestion on stairs and has decided to implement a semaphore-based traffic-control system. Most connections between floors have two flights of stairs with an intermediate landing (see figure). The constraints the Safety Office wishes to enforce are

- Only 1 person at a time on each flight of stairs
- A maximum of 3 persons on a landing
- As a few traffic constraints as possible
- No deadlock (a particular concern if there’s bidirectional travel)

Assume stair traffic is unidirectional: once on a flight of stairs, people continue up or down until they’ve reached their destination floor (no backing up!), although they may pause at the landing.

There are three semaphores: they control the upper flight of stairs (SU), the landing (L), and the lower flight of stairs (SL). Please provide appropriate initial values for these semaphores and add the necessary wait() and signal() calls to the Down() and Up() procedures below. Note that the Down() and Up() routines will be executed by many students simultaneously and the semaphores are the only way their code has of interacting with other instances of the Down() and Up() routines. To get full credit your code must avoid deadlock and enforce the stair and landing occupancy constraints. **Hint:** for half credit, implement a solution where only 1 person at time is in-between floors (but be careful of deadlock here too!).

```plaintext
// Semaphores shared by all students, provide initial values
semaphore SU = __1__, SL = __1__, L = __3__;

// code for going downstairs
Down() {
    wait(L)
    wait(SU)
    Enter SU;
    Exit SU/enter landing;
    signal(SU)
    wait(SL)
    Exit landing/enter SL;
    signal(SL)
    Exit SL;
    signal(SL)
}

// code for going upstairs
Up() {
    wait(L)
    wait(SL)
    Enter SL;
    Exit SL/enter landing;
    signal(SL)
    wait(SU)
    Exit landing/enter SU;
    signal(SL)
    Exit SU;
    signal(SL)
}
```
Problem 6.

(A) Semaphore S is used to implement mutual exclusion on accesses to a shared buffer. No other
semaphores are used. What should its initial value be?

A value of 1 for S allows at most one WAIT to

succeed – others will stall until first one SIGNALs.  

Initial value for S: __________

This implements mutual exclusion.

(B) Indicate whether each of the following sets of semaphore-synchronized processes can
deadlock. The last two cases are variants of the first one; differences are underlined.

<table>
<thead>
<tr>
<th>Initial semaphore values:</th>
<th>P1:</th>
<th>P2:</th>
<th>P3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1 = 1, s2 = 1, s3 = 1</td>
<td>wait(s1);</td>
<td>wait(s2);</td>
<td>wait(s1);</td>
</tr>
<tr>
<td></td>
<td>wait(s2);</td>
<td>wait(s3);</td>
<td>wait(s2);</td>
</tr>
<tr>
<td></td>
<td>print(“1”);</td>
<td>print(“2”);</td>
<td>wait(s3);</td>
</tr>
<tr>
<td></td>
<td>signal(s2);</td>
<td>signal(s3);</td>
<td>print(“3”);</td>
</tr>
<tr>
<td></td>
<td>signal(s1);</td>
<td>signal(s2);</td>
<td>signal(s3);</td>
</tr>
</tbody>
</table>

Can it deadlock?

Uses a global ordering for semaphores (S1 > S2 > S3). No deadlock possible.

Deadlock scenario:
P1 holds S1, waiting on S2
P3 holds S2, waiting on S1

Can it deadlock?

Now both WAIT(S1) statements will succeed.

Can it deadlock?

Circle answers below