CMU 15-418/618: Parallel Computer Architecture and Programming
Practice Exercise 3

Problem 1: Parallel Histogram Generation (Yet Again)

Your friend implements the following parallel code for generating a histogram from the values in a large input array `input`. For each element of the input array, the code uses the function `bin_func` to compute a "bin" the element belongs to (`bin_func` always returns an integer between 0 and `NUM_BINS-1`), and increments a count of elements in that bin. His port targets a small parallel machine with only two processors. *This machine features 64-byte cache lines and uses an invalidation-based cache coherence protocol.* Your friend's implementation is given below.

```
float input[N];                   // assume input is initialized and N is a very large
int   histogram_bins[NUM_BINS];   // output bins
int   partial_bins[2][NUM_BINS];  // assume bins are initialized to 0
                                  // assume partial_bins is 64-byte aligned

////////////////////////// Code executed by thread 0 //////////////////////////
for (int i=0; i<N/2; i++)
   partial_bins[0][bin_func(input[i])]++;

barrier();  // wait for both threads to reach this point

for (int i=0; i<NUM_BINS; i++)
    histogram_bins[i] = partial_bins[0][i] + partial_bins[1][i];

////////////////////////// Code executed by thread 1 //////////////////////////
for (int i=N/2; i<N; i++)
   partial_bins[1][bin_func(input[i])]++;

barrier();  // wait for both threads to reach this point
```

A. (4 pts) Your friend runs this code on an input of 1 million elements (`N`=1,000,000) to create a histogram with eight bins (`NUM_BINS`=8). He is shocked when his program obtains far less than a linear speedup, and glumly asserts believe he needs to completely restructure the code to eliminate load imbalance. You take a look and recommend that he not do any coding at all, and just create a histogram with 16 bins instead. Who's approach will lead to better parallel performance? Why?

B. (4 pts) Inspired by his new-found great performance, your friend concludes that more bins is better. He tries to use the provided code from part A to compute a histogram of 10,000 elements with 2,000 bins. He is shocked when the speedup obtained by the code drops. Improve the existing code to scale near linearly with the larger number of bins. (Please provide pseudocode as part of your answer – it need not be compilable C code.)

C. (2 pts) Your friend changes `bin_func` to a function with *extremely high arithmetic intensity*. (The new function requires 100000's of instructions to compute the output bin for each input element). If the histogram code **provided in part A** is used with this new `bin_func` do you expect scaling to be better, worse, or the same as the scaling you observed using the old `bin_func` in part A? Why? (Please ignore any changes you made to the code in part B for this question.)

Problem 2: Angry Students

Your friend is developing a game that features a horde of angry students chasing after professors for making long exams. Simulating students is expensive, so your friend decides to parallelize the computation using one thread to compute and update the student's positions, and another thread to simulate the student's angriness. The state of the game's N students is stored in the global array `students` in the code below).

```
struct Student {
   float position;   // assume 1D position for simplicity
   float angriness;
};

Student students[N];

////////////////////////////////

void update_positions() {
   for (int i=0 i<N; i++) {
      students[i].position = compute_new_position(i);
   }
}

void update_angriness() {
   for (int i=0 i<N; i++) {
      students[i].angriness = compute_new_angriness(i);
   }
}

////////////////////////////////

// ... initialize students here

pthread_t t0, t1;
pthread_create(&t0, NULL, updatePositions, NULL);
pthread_create(&t1, NULL, updateAngriness, NULL);
pthread_join(t0, NULL);
pthread_join(t1, NULL);
```

Questions are on the next page...

A. (4 pts) Since there is no synchronization between thread 0 and thread 1, your friend expects near a perfect 2× speedup when running on two-core processor that implements invalidation-based cache coherence. She is shocked when she doesn't obtain it. Why is this the case? (For this problem assume that there is sufficient bandwidth to keep two cores busy – "the code is bandwidth bound" is not an answer we are looking for.)

B. (6 pts) Modify the program to correct the performance problem. You are allowed to modify the code and data structures as you wish, **but you are not allowed to change what computations are performed by each thread and your solution should not substantially increase the amount of memory used by the program.** You only need to describe your solution in pseudocode (compilable code is not required).