ECE 454 – Computer Systems Programming

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering

Final Examination Fall 2011

Name	
Student #	

Professor Greg Steffan

Answer all questions. Write your answers on the exam paper. Show your work.
Each question has a different assigned value, as indicated.

The exam is open book open notes (only simple calculators allowed, no cell phones or PDAs)

Total time available: **150 minutes**

Total marks available: **145** (roughly one mark per minute, with 5 extra minutes)

Verify that your exam has all the pages.

Part	Points	Mark
1	30	
2	20	
3	10	
4	10	
5	10	
6	20	
7	10	
8	15	
9	20	
Total	145	

PART 1) [30] Short Answer

a) Modern multicores often have on-chip L2 caches as well as L1 caches. It is common to implement something called the "inclusion property", which means that if a certain cache block is evicted from the L2 cache (eg., due to a conflict) then it also must be evicted from the L1 cache. In one sentence, why would this property be useful?

5 marks
coherence mechanisms don't have to check both the L2 and the L1 for existence of a cache block, only the L2.

b) Suppose that you are planning to parallelize the following loop, where N is very large. However, you are considering first unrolling the loop four times (each new iteration would contain the work of four original iterations).

```
for (i=0;i<N;i++){
    x = work(i);
}
```

i) what are the properties of an instance of work() for which unrolling would likely improve parallel performance (compared to not unrolling), and why?

2marks
if work is regular/predictable, reduce loop management overhead, possibly better cache locality.

ii) what are the properties of an instance work() for which unrolling would likely reduce parallel performance (compared to not unrolling), and why?

2marks
if work is random duration with a wide variance, or some pattern that gives load imabalance for the 4-way unrolling.

c) You have created an implementation of an MCS lock: a linked-list-based queueing lock, where the lock() operation causes a thread to enqueue a lock_entry structure to the end of the lock-queue, and then spin on that lock_entry. The unlock() operation dequeues & frees its lock_entry and notifies the next lock_entry in the queue.

i) Your lock_entry struct contains a lock variable, pointers, and a few fields for tracking statistics. By reordering/optimizing the struct declaration you are able to reduce the size of each lock_entry structure from 64B to 32B. However, you find that performance of the reduced-size version is worse! How can this be the case?

2 marks
If the machine has 64B cache blocks then the 32B version likely has more false sharing.

ii) Assuming you keep the 64B version of lock_entry, should you consider adding "back-off" to the spin-wait implementation for the lock() routine? Why or why not?

2 marks
No, since we are spinning locally back-off wouldn't add much value, and possibly more delay on realizing that it is your turn (although back-off probably saves power).

d) Thread1 executes code that stores 5 to location X then loads X. In parallel thread2 executes code that stores 7 to X. There is no synchronization in either code. When Thread1 loads X it might get the value 7 or might get the value 5. This is because of the processor's implementation of coherence or consistency?

4 marks
coherence

e) It is possible to implement MPI on top of a shared memory multiprocessor or multicore. Suppose you were to implement MPI using pthreads, i.e., using the following pthreads functions as described in class: `pthread_create()`, `pthread_join()`, `pthread_mutex_lock()`, `pthread_mutex_unlock()`, `pthread_cond_signal()`, `pthread_cond_wait()`, `semaphore_signal()`, `semaphore_wait()`.

i) which of the above pthreads functions would you use in your code to implement a non-blocking MPI_Send() and briefly why?

2 marks
semaphore_signal() since you need to signal that your sent data is ready, and you want to "remember the signal".

ii) which of the above pthreads functions would you use in your code to implement a blocking MPI_Recv() and briefly why?

2 marks
semaphore_wait() since you want to block awaiting the signal from the sending thread.

f) Comparing transactional memory and locks

i) In one sentence explain how Transactional Memory is similar to coarse-grain locking.

2marks
TM provides an API that is similar to having a single global lock.

ii) In one sentence explain how Transactional Memory is similar to fine-grain locking.

2marks
In theory TM provides performance similar or possibly better than that of fine-grain locking.

iii) In one sentence explain the behavior of a critical section (executed by many parallel threads) within a loop where it would be better for performance to protect it with a regular lock than with a Transactional Memory transaction.

2marks
A high-contention critical section.

g) Altera visit question:

Some algorithms can produce different solutions to the same problem depending on how long they are allowed to execute. When parallelizing such an algorithm, what do you have to be careful of when deciding whether the parallel version is an improvement?

3 marks
That the quality-of-result (QoR) for the parallel version is actually better than the QoR for the single-thread version if it were executed for the same duration as the parallel version.

PART 2) [20] Cache Optimization

Optimize the cache performance of the following code. Do not parallelize it. Do not consider prefetching nor TLB behavior. You can assume that none of the operations overflow. Assume that N and M are extremely large powers of 2. Assume that A[], B[], andC[] are integer arrays, and that an int is 4B.

i)

```
for (i=0;i<N;i++)
  for (j=0;j<M;j++)
    A[i] += B[j][j] * C[i];
```

6 marks

```
int tmp = 0;
for (j=0;j<M;j++)
  tmp += B[j][j];
for (i=0;i<N;i++)
  A[i] += tmp * C[i];
```

ii)

```
idx = 0;
for (i=0;i<N;i++)
  for (j=0;j<M;j++)
    if (A[i] == B[j])
      C[idx++] = hash(i,j);
```

14 marks

```
idx = 0;
int bsize = 2048;
for (i=0;i<N/bsize;i++)
  for (j=0;j<M/bsize;j++)
    for (i2=i*bsize;i2<(i+1)*bsize;i2++)
      for (j2=j*bsize;j2<(j+1)*bsize;j2++)
        if (A[i2] == B[j2])
          C[idx++] = hash(i2,j2);
```

PART 3) [10] Performance Measurement

Your code goes 1.4x faster after applying 4-thread perfect parallelization to loop1 and 8-wide perfect "SIMDization" (eg., exploiting SSE instructions) to loop2. Knowing that loop1 is 20% of the original (pre-optimization) execution time, what percentage of the original (pre-optimization) program execution time is loop2? You can assume that loop1 and loop2 are not nested.

Speedup = old/new = 1.0/new = 1.4 // can work assuming old time is 1.0

1.0/(1 - 0.2 - x + 0.2/4 + x/8) = 1.4

1/1.4 = 1 - 0.2 - x + 0.2/4 + x/8

x - x/8 = 1 - 0.2 + 0.2/4 - 1/1.4

8*(x - x/8) = (1 - 0.2 + 0.2/4 - 1/1.4)*8

7x = (1 - 0.2 + 0.2/4 - 1/1.4)*8

x = (1 - 0.2 + 0.2/4 - 1/1.4)*8/7

x = 0.155

ANSWER: L2 is 15.5% of execution

PART 4) [10] Prefetching

Rewrite the following code with prefetching inserted for the array `D`. Use the function `pf(int *addr)` to make a prefetch (eg., `pf(&(x))` would prefetch x). Assume that a prefetch takes as long as 32 cache hits (ie., 32 memory accesses that hit in the cache can "hide" the latency of a prefetch). Assume that `D[]` is an array of pointers to arrays, that the cache block size is 32bytes, and that an int is 4bytes. Assume that only accesses to the array generate cache accesses (i.e., ignore accesses to i, j, sum). You should only issue one prefetch for a given cache block.

```
for (i=0;i<N;i++){ // loop1
  for (j=0;j<N;j++){ // loop2
    sum += D[i][j];
  }
}
```

```
for (i=0;i<N;i++){
  for (t1=0;t1<N/8;t1++){
    int *tmpD = D[i];
    pf(&(tmpD[t1+32]))
    for (j=t1*8;j<(t1+1)*8;j++){
      sum += tmpD[j];
    }
  }
}
```

PART 5) [10] TLBs

Suppose your C-code reads and updates a single 64x64 array of 128B structs repeatedly, visiting all of the elements of all of the rows in order. Assuming a 128-entry TLB and 4KB pages, would tiling this code into 4 tiles reduce TLB misses? If so by what fraction? If not, show your calculations as to why not.

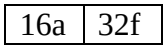
ANSWER: each row of the array consumes 64*128B = (2^6)*(2^7) = 2^13B
each page is 4KB = 2^12B
therefore need 2^13B/2^12B = 2 pages per row
therefore need 2pages-per-row * 64rows = 128pages total
since the TLB is 128 entries, tiling won't reduce TLB misses

PART 6) [20] Dynamic Memory Allocation

Consider an allocator with the following:

- Uses a single implicit free list.
- All memory blocks have a size that is a multiple of 4 bytes and is at least 8 bytes.
- All headers, footers, and pointers are 4 bytes in size
- Allocated blocks consist of a header and a payload (no footer)
- Free blocks consist of a header and a footer at the end of the block.
- All freed blocks are immediately coalesced if possible.
- The heap starts with 0 bytes, never shrinks, and only grows large enough to satisfy memory requests.
- The heap contains only allocated blocks and free blocks. There is no space used for other data or special blocks to mark the beginning and end of the heap.
- When a block is split, the lower part of the block becomes the allocated part and the upper part becomes the new free block (and is coalesced if possible).
- The **best-fit** free block is used when searching the free list.
- If no free block is large enough then the heap is extended only enough to fulfill the request, including using any remaining free block at the end of the free list.

Below you are given a series of memory requests. You are asked to show what the heap looks like after each request is completed using the best-fit placement policy. The heap is represented as a series of boxes, where each box is a single block on the heap, and the bottom of the heap is the left most box. In each block, you should write the total size (including headers and footers) of the block in bytes and either 'f' or 'a' to mark it as free or allocated, respectively. For example, the following heap contains an allocated block of size 16, followed by a free block of size 32.



Assume that the heap is empty before each of the sequences is run, with a single 400B free block (shown for you). You do not necessarily have to use all of the columns provided for the heap. There are three copies here in case you make a mess. Clearly cross out the ones you don't want graded.

COPY #1: (they are identical, cross out the ones you don't want graded)

	400f						
ptr1 = malloc(8)	12a	388f					
ptr2 = malloc(31)	12a	36a	352f				
free(ptr1)	12f	36a	352f				
ptr3 = malloc(29)	12f	36a	36a	316f			
ptr4 = malloc(4)	12f	36a	36a	8a	308f		
ptr5 = malloc(312)	12f	36a	36a	8a	316a		
free(ptr4)	12f	36a	36a	8f	316a		
ptr6 = malloc(9)	12f	36a	36a	8f	316a	16a	
ptr7 = malloc(4)	12f	36a	36a	8a	316a	16a	
free(ptr3)	12f	36a	36f	8a	316a	16a	
free(ptr2)	84f	8a	316a	16a			
free(ptr6)	84f	8a	316a	16f			

COPY #2: (they are identical, cross out the ones you don't want graded)

	400f						
ptr1 = malloc(8)							
ptr2 = malloc(31)							
free(ptr1)							
ptr3 = malloc(29)							
ptr4 = malloc(4)							
ptr5 = malloc(312)							
free(ptr4)							
ptr6 = malloc(9)							
ptr7 = malloc(4)							
free(ptr3)							
free(ptr2)							
free(ptr6)							

COPY #3: (they are identical, cross out the ones you don't want graded)

	400f						
ptr1 = malloc(8)							
ptr2 = malloc(31)							
free(ptr1)							
ptr3 = malloc(29)							
ptr4 = malloc(4)							
ptr5 = malloc(312)							
free(ptr4)							
ptr6 = malloc(9)							
ptr7 = malloc(4)							
free(ptr3)							
free(ptr2)							
free(ptr6)							

PART 7) [10] Cache Coherence

Assuming a 4-CPU multicore with MESI invalidation-based cache coherence with write-allocate write-back caches, and locations X and Y that are in the same cache block, in the table below are shown the order in time of loads and stores to X and Y. Put a '1' or a checkmark in the column on the left next to each row for which a coherence message occurs that requests a copy of the cache block contents for the corresponding CPU's cache. HINT: a load-miss results in a copy request, while a write-hit does not.

Copy of cache block requested?	CPU 0	CPU 1	CPU 2	CPU 3
1		Store X		
		Load X		
1			Store Y	
			Load X	
1				Load X
1	Load X			
1		Load X		
			Store Y	
1		Store X		
1			Load Y	
		Load X		
		Load Y		
1				Load X
				Store Y

Minus two marks for any checkmark missing/extra

PART 8) [15] Dependence Analysis

For each loop in the following code, state whether or not it is parallel, and if not describe the dependence(s) that prevent it from being parallel (give the type of dependence and which terms c1,c2,c3,c4 are involved).

```
for (i=2;i<N-1;i++){ // loop1
  for (j=3;j<N;j++){ // loop2
    for (k=4;k<N-3;k++){ // loop3
      A[i][j][k] =                      // c1
                   A[i-2][j-3][k-4] + // c2
                   A[i+1][j][k+2] +   // c3
                   A[i][j][k-1];      // c4
    }
  }
}
```

loop1: not parallel, 2marks

flow dep C1->C2, 3marks

anti dep C1->C3 3marks

loop2: parallel 2marks

loop3: not parallel, 2marks

flow dep C1->C4 3marks

PART 9) [20] OpenMP Parallelization

Parallelize the following code using openMP pragmas. Assume that the target machine has a cache block size of 128B, and that the size of an int is 4B. Be sure to explicitly specify the "schedule" options that should be used, even if you want to use the default options. For each please rewrite the code. If necessary you can assume that the variable P represents the number of processors to be used. Assume that N is large (in the tens of thousands or more). You must explicitly list all variables within the range of a parallel pragma that are private using the private() directive. Do not do tiling. Please rewrite the new code.

a)

```
for (i=0;i<N;i++){
    for (j=0;j<N;j++){
        A[i,j] = max(A[i,j],B[i,j]);
    }
}
```

5marks

```
#pragma omp parallel for private(j) schedule(static,N/P)
for (i=0;i<N;i++){
    for (j=0;j<N;j++){
        A[i,j] = max(A[i,j],B[i,j]);
    }
}
```

b)

```
C[0] = 1;
for (i=1;i<N;i++){
    C[i] = C[i-1];
    for (j=0;j<N;j++){
        C[i] *= A[i,j] + B[i,j];
    }
}
```

5marks

```
C[i] = 1;
for (i=1;i<N;i++){
    C[i] = C[i-1];
#pragma omp parallel for schedule(static,N/P)
reduction(*:product)
    for (j=0;j<N;j++){
        product *= A[i,j] + B[i,j];
    }
    C[i] = product;
}
```

c)

```
typedef struct element
{
  int value;
  struct element *next;
} Element;

Element *D[N]; // D[] is an array of pointers to linked lists
of varying length
int C[N];
...
for (i=0;i<N;i++){
    C[i] = computeAverageValueOfAllListElements(D[i]);
}
```

5marks

```
#pragma omp parallel for schedule(dynamic,1)
for (i=0;i<N;i++){
    C[i] = computeAverageValueOfAllListElements(D[i]);
}
```

d)

```
// Assume the same D[] array from part (c)
Element *pointer_to_max = NULL;
for (i=0;i<N;i++){
    Element *tmpPtr = findListElementWithMaxValue(D[i]);
    if (!pointer_to_max ||
        (tmpPtr && tmpPtr->value > pointer_to_max->value))
       pointer_to_max = tmpPtr;
}
```

5marks

```
#pragma omp parallel for private(tmpPtr) schedule(dynamic,1)
for (i=0;i<N;i++){
    Element *tmpPtr = findListElementWithMaxValue(D[i]);
    #pragma omp critical
    if (!pointer_to_max_element ||
        (tmpPtr && tmpPtr->value > pointer_to_max->value))
       pointer_to_max = tmpPtr;
}
```