

## 1 Pre-Check

This section is designed as a conceptual check for you to determine if you conceptually understand and have any misconceptions about this topic. Please answer true/false to the following questions, and include an explanation:

- 1.1 Both the multithreading in data-level parallelism and the manager-worker framework used in multiprocess code do not use shared memory.

False. Multithreaded programs can access main memory across threads, causing data races if written incorrectly. On the other hand, multiprocess programs have completely independent and distinct instances of the program starting from `MPI_Init`.

- 1.2 Because the manager-worker framework requires one process to deal with load balancing the rest of the work across programs, process-level parallelism is mostly useful for large-scale tasks.

True. Open MPI requires massive amounts of overhead, moreso than any other form of parallelism discussed in this course, with an entire dedicated manager process and the expensive communication across individual nodes.

- 1.3 Because process-level parallelism already takes advantage of multiple nodes, utilizing the OpenMP library in the Open MPI framework results in a performance decrease, as each thread will do the same, redundant work.

False. Thread-level parallelism does its multi-threaded work onto one node, as all its work is done onto one shared memory, while process-level parallelism can work across nodes. While both forms of parallelism allow for multiple operations to be done concurrently, the resources each require and can use are different. If allocated correctly, OpenMp and Open MPI can end up being complementary to each other, and are necessary optimizations in supercomputers, where much more resources are available and operations are done on a massive scale.

## 2 Open MPI

Beyond multithreading, the idea of process-level programming is to run one program on multiple processes at once.

The Open MPI project provides a way of writing programs which can be run on multiple processes. We can use its C libraries by calling their functions. Then, when we run the program, Open MPI will create a bunch of processes and run a copy of the code on each process. Here is a list of the most important functions for this class:

- **int MPI\_Init(int\* argc, char\*\*\* argv)** should be called at the start of the program, passing in the addresses of argc and argv.
- **int MPI\_Finalize()** should be called at the end of the program.
- **int MPI\_Comm\_size(MPI\_COMM\_WORLD, int \*size)** gets the total number of processes running the program, and puts it in size.
- **int MPI\_Comm\_rank(MPI\_COMM\_WORLD, int \*rank)** gets the ID of the current process (0 ~ total number of processes - 1) and puts it in rank.
- **int MPI\_Send(const void \*buf, int count, MPI\_Datatype datatype, int dest, 0, MPI\_COMM\_WORLD)** sends a message in buf, which consists of count things with data type datatype to the process with ID dest.
- **int MPI\_Recv(void \*buf, int count, MPI\_Datatype datatype, int source, 0, MPI\_COMM\_WORLD, MPI\_Status \*status)** receives a message consisting of count things with data type datatype from the process with ID source, and puts the message into buf. Some additional information is put into a struct at status.
  - If you want to receive a message from any source, set the source to be MPI\_ANY\_SOURCE.
  - The source of the message can be found in the MPI\_SOURCE field of the outputted status struct.
  - If you don't need the information in the status struct (e.g. because you already know the source of the message), set the status address to MPI\_STATUS\_IGNORE.

**Note:** Unlike OpenMP, the MPI functions will always put their results into an address which you provide as their arguments. The return value of the function is not an output, but rather the error code of the function. In this section, we will implement the ManyMatMul example from lecture using a manager-worker approach.

We have  $n$  pairs of matrices available in input files `Task0a.mat`, `Task0b.mat`, `Task1a.mat`, `Task1b.mat`, ..., and we want to multiply each pair of matrices together, with their outputs written to the output files `Task0ab.mat`, `Task1ab.mat`, ...

We want to accomplish this task using multiple processes such that one process (the manager) assigns work to all other available processes (the workers).

- 2.1 First, perform the overall setup required for Open MPI to function. Fill out the following skeleton of the program:

```

1  #define TERMINATE -1
2  #define READY 0
3
4  /**
5   * Takes in a number i. Reads files Taskia.mat, Taskib.mat,
6   * multiplies them, then outputs to Taskiab.mat.
7   */
8  int matmul(int i) {
9      // omitted
10 }
11
12 int main(int argc, char** argv) {
13     int numTasks = atoi(argv[1]); // read n from command line
14     MPI_Init(&argc, &argv); // initialize
15     // get process ID of this process and total number of processes
16     int procID, totalProcs;
17     MPI_Comm_size(MPI_COMM_WORLD, &totalProcs);
18     MPI_Comm_rank(MPI_COMM_WORLD, &procID);
19     // are we a manager or a worker?
20     if (procID == 0) {
21         // manager node code (see Q2.3)
22     } else {
23         // worker node code (see Q2.2)
24     }
25     MPI_Finalize(); // clean up
26 }

```

- 2.2 Next, fill in what the worker needs to do. Worker processes should repeatedly ask the manager for more work, then perform the work the manager asks of it. If it receives a message that there's no work to be done, it should stop. Let us define a simple communication protocol between the manager and worker:

- When the worker is free, it will send the READY(0) message to the manager.
- The manager will send one number back, which is the task number the worker should work on next.
- If there are no more tasks to done, then instead the manager will send back the TERMINATE(-1) message to the worker.

We will use a single 32-bit signed integer as the message, which corresponds to the MPI data type MPI\_INT32\_T.

```

1 // worker node code
2 int32_t message;
3 while (true) {
4     // request more work
5     message = READY;
6     MPI_Send(&message, 1, MPI_INT32_T, 0, 0, MPI_COMM_WORLD);
7     // receive message from manager
8     MPI_Recv(&message, 1, MPI_INT32_T, 0, 0, MPI_COMM_WORLD, MPI_STATUS_IGNORE);
9     if (message == TERMINATE) break; // all done!
10    matmul(message); // do work
11 }

```

2.3 Finally, fill in the code for the manager process. While there's still more work to do, the manager should wait for a message from any worker and respond with the next task for the worker to work on. When all work has been allocated, the manager should wait for another message from each worker (meaning the worker is done with all work), and respond to each with the TERMINATE(-1) message. The manager shouldn't exit before sending TERMINATE to every worker!

```

1 // manager node code
2 int nextTask = 0; // next task to do
3 MPI_Status status;
4 int32_t message;
5 // assign tasks
6 while (nextTask < numTasks) {
7     // wait for a message from any worker
8     MPI_Recv(&message, 1, MPI_INT32_T, MPI_ANY_SOURCE, 0, MPI_COMM_WORLD, &status);
9     int sourceProc = status.MPI_SOURCE; // process ID of the source of the message
10    // assign next task
11    message = nextTask;
12    MPI_Send(&message, 1, MPI_INT32_T, sourceProc, 0, MPI_COMM_WORLD);
13    nextTask++;
14 }
15 // wait for all processes to finish
16 for (int i = 0; i < totalProcs - 1; i++) {
17     // wait for a message from any worker
18     MPI_Recv(&message, 1, MPI_INT32_T, MPI_ANY_SOURCE, 0, MPI_COMM_WORLD, &status);
19     int sourceProc = status.MPI_SOURCE; // process ID of the source of the message
20     message = TERMINATE;
21     MPI_Send(&message, 1, MPI_INT32_T, sourceProc, 0, MPI_COMM_WORLD);
22 }

```

### 3 Open MPI with Dependencies

Now that we have a working Open MPI implementation of our ManyMatMul task, lets extend this to account for data dependencies! Let's change our task to have an additional step: multiply  $n$  output matrices `Task0ab.mat`, `Task1ab.mat`, etc. in place with a set matrix `kernel.mat`.

Here we provide a new function to use in the worker process:

```

1  /**
2   * Takes in a number i. Reads files Taskiab.mat and
3   * multiplies them with kernel.mat in place. If file
4   * does not exist, return -1
5   */
6  int final_matmul(int i) {
7      //omitted
8  }
```

- 3.1 Provided below is the pseudocode for the manager process in our new implementation. Assume that our program and workers are set up in the same way as described in Q3.

```

1  // manager node pseudocode
2  counter = 0;
3  while (counter < n) {
4      Wait for a message from any worker;
5      Assign worker with the next pair of matrices to multiply,
6      worker will call matmul(counter);
7      counter++;
8  }
9  counter = 0;          // start in-place multiplication
10 while (counter < n) {
11     Wait for a message from any worker;
12     Assign worker with next in-place multiplication,
13     worker will call final_matmul(counter);
14     counter++;
15 }
16 // wait for all processes to finish
17 for each process {
18     Wait for a message from any worker;
19     Send worker message to TERMINATE;
20 }
```

Will this program successfully output the correct matrix files? If it doesn't, explain why. If it does, does it optimally parallelize our desired task? You may assume that if `final_matmul` returns -1, the worker will wait some amount of time before sending the manager another `READY` message.

As the second while loop does its work in sequential order, the program will be forced to wait for the corresponding first task to finish before attempting any additional

`final_matmults`. For example, if `Task1` was a massive, high-dimensional calculation, each other process would need to wait for the `Task1` to finish before attempting any of the in-place multiplications in the second while loop, creating a performance bottleneck.

## 4 Critical Sections

- 4.1 Consider the following multithreaded code to compute the product over all elements of an array.

```

1 // Assume arr has length 8*n.
2 double fast_product(double *arr, int n) {
3     double product = 1;
4     #pragma omp parallel for
5     for (int i = 0; i < n; i++) {
6         double subproduct = arr[i*8]*arr[i*8+1]*arr[i*8+2]*arr[i*8+3]
7             * arr[i*8+4]*arr[i*8+5]*arr[i*8+6]*arr[i*8+7];
8         product *= subproduct;
9     }
10    return product;
11 }
```

- (a) What is wrong with this code?

The code has the shared variable `product`, which can cause data races when multiple threads access it simultaneously.

- (b) Fix the code using `#pragma omp critical`. What line would you place the directive on to create that critical section?

```

1 double fast_product(double *arr, int n) {
2     double product = 1;
3     #pragma omp parallel for
4     for (int i = 0; i < n; i++) {
5         double subproduct = arr[i*8]*arr[i*8+1]*arr[i*8+2]*arr[i*8+3]
6             * arr[i*8+4]*arr[i*8+5]*arr[i*8+6]*arr[i*8+7];
7         #pragma omp critical
8         product *= subproduct;
9     }
10    return product;
11 }
```

- 4.2 When added to a `#pragma omp parallel for` statement, the `reduction(operation : var)` directive creates and optimizes the critical section for a for loop, given a variable that should be in the critical section and the operation being performed on that variable. An example is given below.

```

1 // Assume arr has length n
2 int fast_sum(int *arr, int n) {
3     int result = 0;
```

```

4     #pragma omp parallel for reduction(+: result)
5     for (int i = 0; i < n; i++) {
6         result += arr[i];
7     }
8     return result;
9 }

```

Fix the code by adding the `reduction(operation: var)` directive to the `#pragma omp parallel for` statement. Which variable should be in the critical section, and what is the operation being performed?

```

1 double fast_product(double *arr, int n) {
2     double product = 1;
3     #pragma omp parallel for reduction (*:product)
4     for (int i = 0; i < n; i++) {
5         double subproduct = arr[i*8]*arr[i*8+1]*arr[i*8+2]*arr[i*8+3]
6             * arr[i*8+4]*arr[i*8+5]*arr[i*8+6]*arr[i*8+7];
7         product *= subproduct;
8     }
9     return product;
10 }

```