CSCI 455: Lab #4 — MPI Timing Model and Comparison

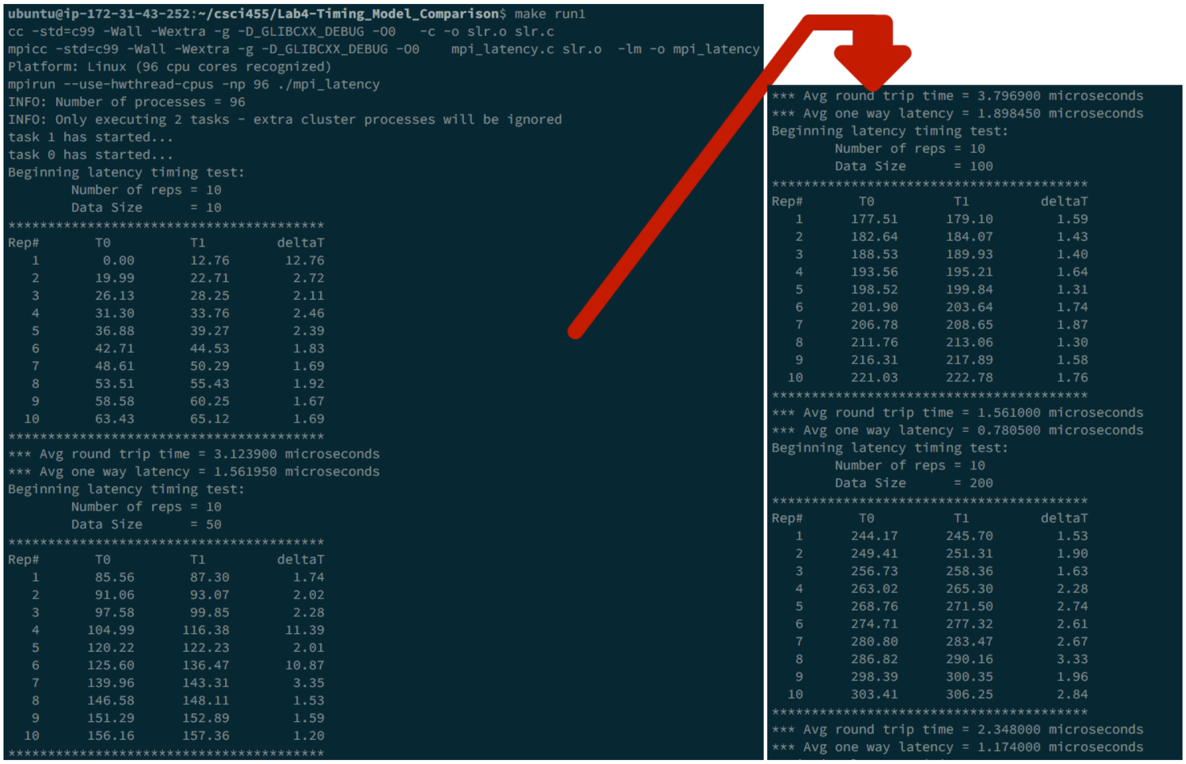
Darwin Jacob Groskleg

Winter 2020

Contents

Summary Lab 4 consists of a series of tests to measure the timing of a single send/recv communication using the ping-pong method. Based on the tests, you can estimate the `t_startup` and `t_data` with the least square regression method.

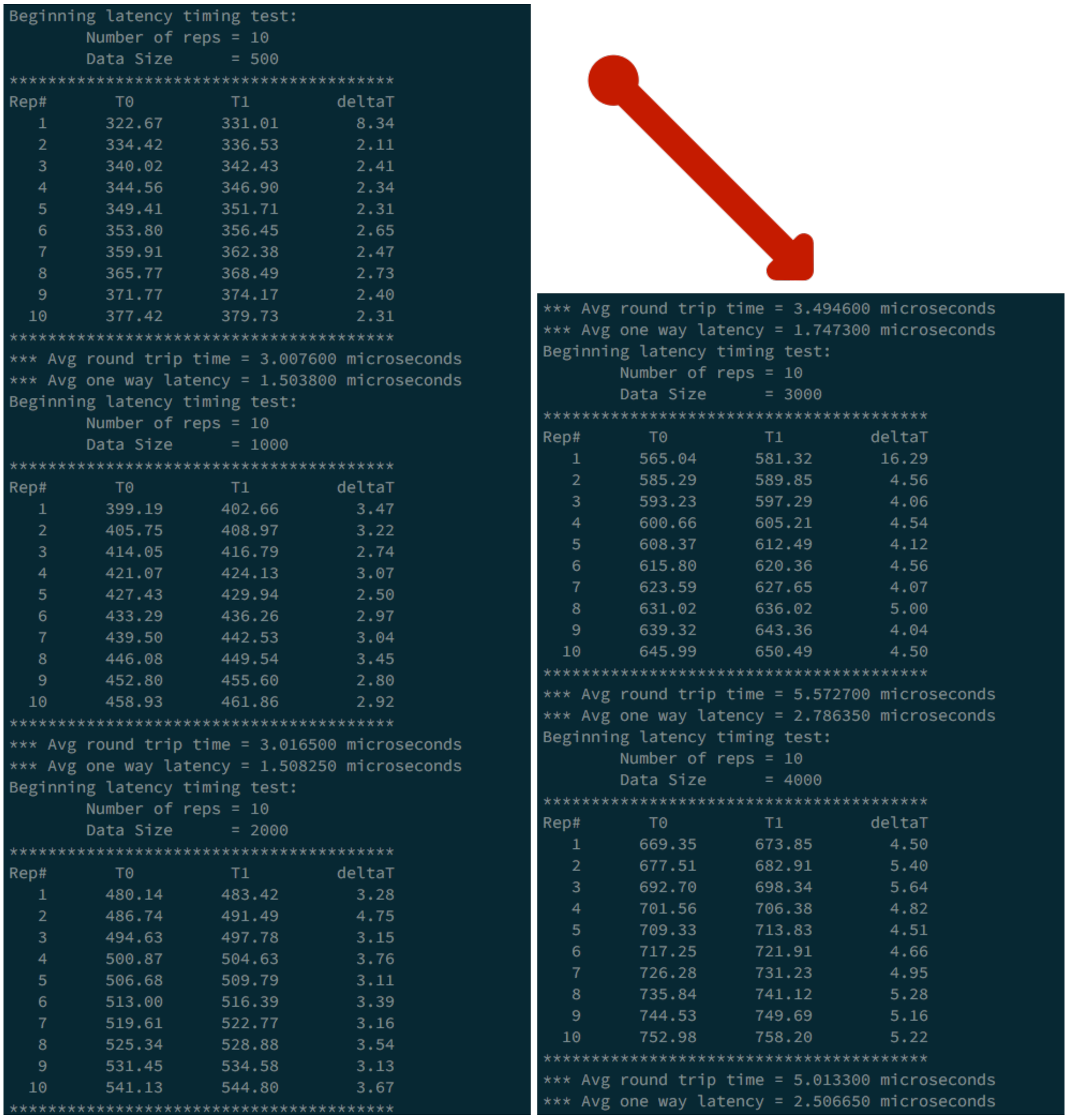
Part 1: Estimating Communication Latency of MPI Send/Recv



```
ubuntu@ip-172-31-43-252:~/csci455/Lab4-Timing_Model_Comparison$ make run1
cc -std=c99 -Wall -Wextra -g -D_GLIBCXX_DEBUG -O0   -c -o slr.o slr.c
mpicc -std=c99 -Wall -Wextra -g -D_GLIBCXX_DEBUG -O0    mpi_latency.c slr.o  -lm -o mpi_latency
Platform: Linux (96 cpu cores recognized)
mpirun --use-hwthread-cpus -np 96 ./mpi_latency
INFO: Number of processes = 96
INFO: Only executing 2 tasks - extra cluster processes will be ignored
task 1 has started...
task 0 has started...
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 10
****************************************
Rep#       T0          T1          deltaT
   1        0.00       12.76       12.76
   2       19.99       22.71        2.72
   3       26.13       28.25        2.11
   4       31.30       33.76        2.46
   5       36.88       39.27        2.39
   6       42.71       44.53        1.83
   7       48.61       50.29        1.69
   8       53.51       55.43        1.92
   9       58.58       60.25        1.67
  10       63.43       65.12        1.69
****************************************
*** Avg round trip time = 3.123900 microseconds
*** Avg one way latency = 1.561950 microseconds
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 50
****************************************
Rep#       T0          T1          deltaT
   1       85.56       87.30        1.74
   2       91.06       93.07        2.02
   3       97.58       99.85        2.28
   4      104.99      116.38       11.39
   5      120.22      122.23        2.01
   6      125.60      136.47       10.87
   7      139.96      143.31        3.35
   8      146.58      148.11        1.53
   9      151.29      152.89        1.59
  10      156.16      157.36        1.20
****************************************
```

```
*** Avg round trip time = 3.796900 microseconds
*** Avg one way latency = 1.898450 microseconds
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 100
****************************************
Rep#       T0          T1          deltaT
   1      177.51      179.10        1.59
   2      182.64      184.07        1.43
   3      188.53      189.93        1.40
   4      193.56      195.21        1.64
   5      198.52      199.84        1.31
   6      201.90      203.64        1.74
   7      206.78      208.65        1.87
   8      211.76      213.06        1.30
   9      216.31      217.89        1.58
  10      221.03      222.78        1.76
****************************************
*** Avg round trip time = 1.561000 microseconds
*** Avg one way latency = 0.780500 microseconds
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 200
****************************************
Rep#       T0          T1          deltaT
   1      244.17      245.70        1.53
   2      249.41      251.31        1.90
   3      256.73      258.36        1.63
   4      263.02      265.30        2.28
   5      268.76      271.50        2.74
   6      274.71      277.32        2.61
   7      280.80      283.47        2.67
   8      286.82      290.16        3.33
   9      298.39      300.35        1.96
  10      303.41      306.25        2.84
****************************************
*** Avg round trip time = 2.348000 microseconds
*** Avg one way latency = 1.174000 microseconds
```

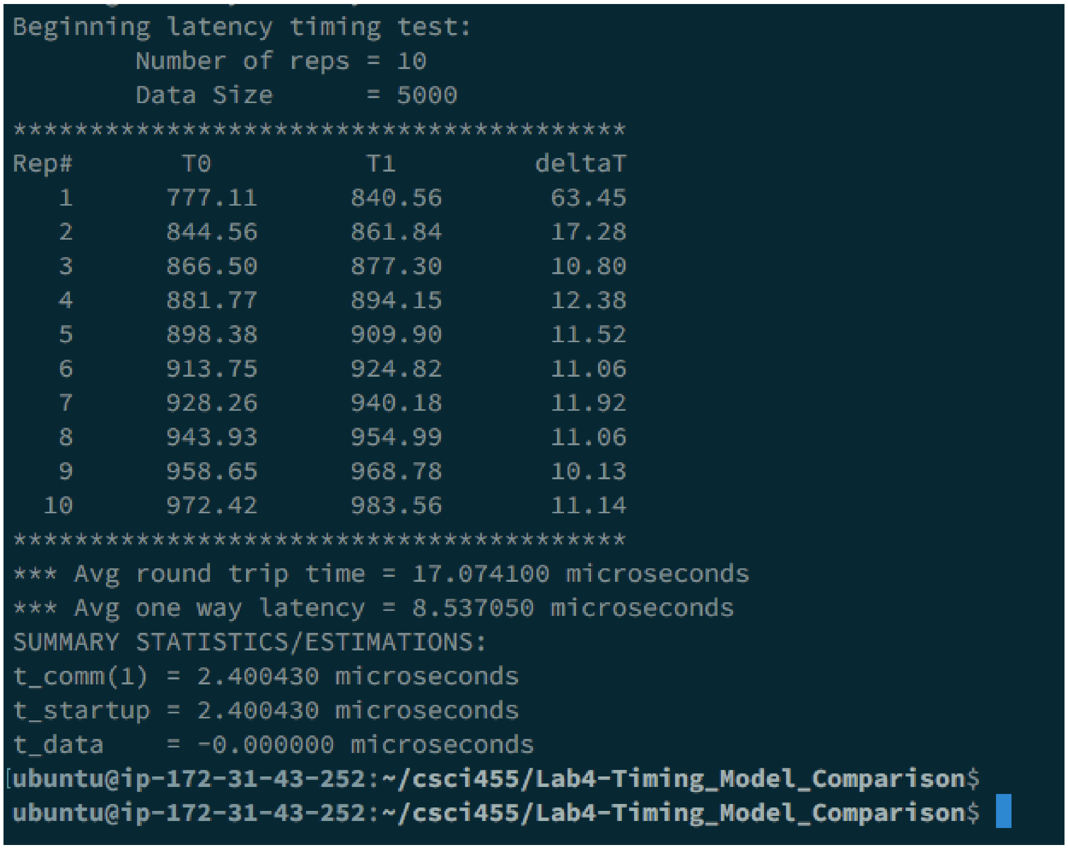
Figure 1: Console screenshot (1/3) of `mpi_latency` running with 96 compute nodes



```
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 500
****************************************
Rep#        T0          T1         deltaT
   1      322.67      331.01         8.34
   2      334.42      336.53         2.11
   3      340.02      342.43         2.41
   4      344.56      346.90         2.34
   5      349.41      351.71         2.31
   6      353.80      356.45         2.65
   7      359.91      362.38         2.47
   8      365.77      368.49         2.73
   9      371.77      374.17         2.40
  10      377.42      379.73         2.31
****************************************
*** Avg round trip time = 3.007600 microseconds
*** Avg one way latency = 1.503800 microseconds
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 1000
****************************************
Rep#        T0          T1         deltaT
   1      399.19      402.66         3.47
   2      405.75      408.97         3.22
   3      414.05      416.79         2.74
   4      421.07      424.13         3.07
   5      427.43      429.94         2.50
   6      433.29      436.26         2.97
   7      439.50      442.53         3.04
   8      446.08      449.54         3.45
   9      452.80      455.60         2.80
  10      458.93      461.86         2.92
****************************************
*** Avg round trip time = 3.016500 microseconds
*** Avg one way latency = 1.508250 microseconds
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 2000
****************************************
Rep#        T0          T1         deltaT
   1      480.14      483.42         3.28
   2      486.74      491.49         4.75
   3      494.63      497.78         3.15
   4      500.87      504.63         3.76
   5      506.68      509.79         3.11
   6      513.00      516.39         3.39
   7      519.61      522.77         3.16
   8      525.34      528.88         3.54
   9      531.45      534.58         3.13
  10      541.13      544.80         3.67
****************************************
```

```
*** Avg round trip time = 3.494600 microseconds
*** Avg one way latency = 1.747300 microseconds
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 3000
****************************************
Rep#        T0          T1         deltaT
   1      565.04      581.32        16.29
   2      585.29      589.85         4.56
   3      593.23      597.29         4.06
   4      600.66      605.21         4.54
   5      608.37      612.49         4.12
   6      615.80      620.36         4.56
   7      623.59      627.65         4.07
   8      631.02      636.02         5.00
   9      639.32      643.36         4.04
  10      645.99      650.49         4.50
****************************************
*** Avg round trip time = 5.572700 microseconds
*** Avg one way latency = 2.786350 microseconds
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 4000
****************************************
Rep#        T0          T1         deltaT
   1      669.35      673.85         4.50
   2      677.51      682.91         5.40
   3      692.70      698.34         5.64
   4      701.56      706.38         4.82
   5      709.33      713.83         4.51
   6      717.25      721.91         4.66
   7      726.28      731.23         4.95
   8      735.84      741.12         5.28
   9      744.53      749.69         5.16
  10      752.98      758.20         5.22
****************************************
*** Avg round trip time = 5.013300 microseconds
*** Avg one way latency = 2.506650 microseconds
```

Figure 2: Console screenshot (1/3) of `mpi_latency` running with 96 compute nodes



```
Beginning latency timing test:
        Number of reps = 10
        Data Size      = 5000
****************************************
Rep#       T0          T1          deltaT
   1      777.11      840.56        63.45
   2      844.56      861.84        17.28
   3      866.50      877.30        10.80
   4      881.77      894.15        12.38
   5      898.38      909.90        11.52
   6      913.75      924.82        11.06
   7      928.26      940.18        11.92
   8      943.93      954.99        11.06
   9      958.65      968.78        10.13
  10      972.42      983.56        11.14
****************************************
*** Avg round trip time = 17.074100 microseconds
*** Avg one way latency = 8.537050 microseconds
SUMMARY STATISTICS/ESTIMATIONS:
t_comm(1) = 2.400430 microseconds
t_startup = 2.400430 microseconds
t_data    = -0.000000 microseconds
ubuntu@ip-172-31-43-252:~/csci455/Lab4-Timing_Model_Comparison$
ubuntu@ip-172-31-43-252:~/csci455/Lab4-Timing_Model_Comparison$
```

Figure 3: Console screenshot (1/3) of `mpi_latency` running with 96 compute nodes. Resulting estimations are `t_startup`=2.4 usec and `t_data`=0.0 usec.

mpi_latency.c

```c
/******************************************************************************
* FILE: mpi_latency.c
* AUTHORS:
*   Darwin Jacob Groskleg (2020)
*   Sazzad (02/11/18)
*   Laurence T. Yang
* DESCRIPTION:
*   MPI Latency Timing Program - C Version
*   In this example code, a MPI communication timing test is performed.
*   MPI process 0 will send "reps" number of 1 byte messages to MPI process 1,
*   waiting for a reply between each rep. Before and after timings are made
*   for each rep and an average calculated when completed.
*
* NOTES
*   - Ping-pong method sends same size data back and forth.
******************************************************************************/

#include <stdio.h>
#include <stdlib.h>
#include <time.h>
#include <string.h>
#include <stdbool.h>
#include <math.h>
#include <assert.h>

#include <sys/time.h>
#include <mpi.h>

#include "slr.h"

#define NUMBER_REPS 10
#define DATA_SIZE 5000

/* send/receive process designators, maps task label to expected rank */
enum TaskRanks {
    Master = 0,
    Worker = 1
};

double sample_latency_with_n_bytes(int bytes_of_traffic);
void run_sampling_responder(int bytes_of_traffic);

/*
 * Estimate the t_startup and t_data with the least square regression method:
 *  y: AvgT/2 (one way latency)
 *  let K = 10 (number of runs, different message sizes)
 * t_startup = m * 0 + b            (time to send msg with no data)
 * t_comm = m*1 + b                 (time to startup and send 1 data word)
 * t_data = t_comm - t_startup   (time to send one data word)
 */
int main (int argc, char *argv[]) {
    MPI_Init(&argc, &argv);
    int cluster_size;           /* number of MPI processes */
    MPI_Comm_size(MPI_COMM_WORLD, &cluster_size);
    int rank;                   /* my MPI process number */
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);

    if (rank == Master && cluster_size != 2) {
        printf(”INFO: Number of processes = %d\n”, cluster_size);
        printf(”INFO: Only executing 2 tasks - extra cluster processes ”);
        printf(”will be ignored\n”);
    }
    MPI_Barrier(MPI_COMM_WORLD);

    if (rank < 2) {
        printf(”task %d has started...\n”, rank);
    }
    MPI_Barrier(MPI_COMM_WORLD);

    int x_byte_counts[] = {10, 50, 100, 200, 500, 1000, 2000, 3000, 4000, 5000};
    size_t k_tests = sizeof(x_byte_counts)/sizeof(x_byte_counts[0]);
    double y_timings[k_tests];

    for (size_t i=0; i<k_tests; i++){
        if (rank == Worker) {
            run_sampling_responder(x_byte_counts[i]);
        }
        else if (rank == Master) {
            y_timings[i] = sample_latency_with_n_bytes(x_byte_counts[i]);
        }
    }

    if (rank == Master) {
        slr_equation_t eqt = slr_find_line(k_tests, x_byte_counts, y_timings);
        double t_startup = slr_predict(eqt, 0);
        double t_comm    = slr_predict(eqt, 1);
        double t_data    = t_comm - t_startup;
        printf(”SUMMARY STATISTICS/ESTIMATIONS:\n”);
        printf(”t_comm(1) = %f microseconds\n”, t_comm);
        printf(”t_startup = %f microseconds\n”, t_startup);
        printf(”t_data    = %f microseconds\n”, t_data);
    }

    MPI_Finalize();
    exit(0);
}

/* Returns: time in microseconds (t_comm)
 *      - the one-way latency (round-trip / 2)
 * Takes:
 *  bytes_of_traffic = N, the amount of data to be sent
 */
double sample_latency_with_n_bytes(int bytes_of_traffic) {
    assert(bytes_of_traffic <= DATA_SIZE);

    char msg[DATA_SIZE];        /* buffer containing DATA_SIZE byte message */
    MPI_Status status;          /* MPI receive routine parameter */
    int tag = 1;                /* MPI message tag parameter */
    int reps = NUMBER_REPS;     /* number of samples per test */
    /* round-trip latency timing test */
    printf(”Beginning latency timing test:\n”);
    printf(”\tNumber of reps = %d\n”, reps);
    printf(”\tData Size      = %d\n”, bytes_of_traffic);
    printf(”****************************************\n”);
    /*      <  >  <      . >  <      . >  <      . >                      */
    printf(”Rep#       T0          T1         deltaT\n”);

    const double MSPerSecond = pow(10, 6);
    const double ClockResolution = 1;//MPI_Wtick(); // seconds per clock tick
    double T0, T1;              /* start/end times per rep in ms */
    double deltaT;              /* time for one rep in ms */
    double sumT = 0;            /* sum of all reps times in microseconds */
    int error;
    for (int n = 1; n <= reps; n++) {
        /* start time */
        T0 = MPI_Wtime() * MSPerSecond * ClockResolution;

        /* send message to worker - message tag set to 1.  */
        error = MPI_Send(
                &msg,
                bytes_of_traffic,
                MPI_BYTE,   // for char
                Worker,     // destination
                tag,
                MPI_COMM_WORLD
            );
        /* If return code indicates error quit */
        // DARWIN GROSKLEG:
        //  this is unecessary and already properly handled
        //  by the MPI error handler, which will abort the MPI job in most
        //  cases.
        if (error)
            MPI_Abort(MPI_COMM_WORLD, error);

        /* Now wait to receive the echo reply from the worker  */
        // ”echoes” the same data back???
        error = MPI_Recv(
            &msg,
            bytes_of_traffic,
            MPI_BYTE,   // for char
            Worker,     // source
            tag,
            MPI_COMM_WORLD,
            &status);

        /* If return code indicates error quit */
        // Redundant step is skipped because the MPI already handles it.
        if (error)
            MPI_Abort(MPI_COMM_WORLD, error);

        /* end time */
        T1 = MPI_Wtime() * MSPerSecond * ClockResolution;

        /* calculate round trip time and print */
        deltaT = T1 - T0;
        sumT += deltaT;
        /* print statement for each to keep each column right justified */
        printf(”%4d  ”, n);
        printf(”%10.2f  ”, T0);
        printf(”%10.2f  ”, T1);
        printf(”%10.2f\n”, deltaT);
    }

    /* average time per rep in microseconds */
    double avgT = sumT / reps;
    printf(”****************************************\n”);
    printf(”*** Avg round trip time = %f microseconds\n”, avgT);
    printf(”*** Avg one way latency = %f microseconds\n”, avgT/2);

    return avgT/2;
}

void run_sampling_responder(int bytes_of_traffic) {
    assert(bytes_of_traffic <= DATA_SIZE);
    char msg[DATA_SIZE];        /* buffer containing DATA_SIZE byte message */
    MPI_Status status;          /* MPI receive routine parameter */
    int tag = 1;                /* MPI message tag parameter */
    int reps = NUMBER_REPS;     /* number of samples per test */

    while (reps--) {
        // ping
        MPI_Recv(
            &msg,
            bytes_of_traffic,
            MPI_BYTE,
            Master,
            tag,
            MPI_COMM_WORLD,
            &status);
        // pong
        MPI_Send(
            &msg,
            bytes_of_traffic,
            MPI_BYTE,
            Master,
            tag,
            MPI_COMM_WORLD);
    }
}
```

slr.h

```
/* slr.h
 * -----
 * Authors: Darwin Jacob Groskleg (2020)
 *
 * Purpose: do simple linear regression.
 */
#ifndef SLR_H_INCLUDED
#define SLR_H_INCLUDED

typedef struct {
    double a; /* intercept */
    double b; /* slope */
} slr_equation_t;

slr_equation_t slr_find_line(int n, int X[], double Y[]);

double slr_predict(slr_equation_t eqt, int x);

#endif /* SLR_H_INCLUDED */
```

slr.c

```c
/* slr.c
 * -----
 * Authors: Darwin Jacob Groskleg (2020)
 *
 * Purpose: do simple linear regression.
 */
#include ”slr.h”

#include <assert.h>
#include <math.h>

double regression_coefficient(int, double, double, double, double);
double y_intercept(int, double, double, double);


double slr_predict(slr_equation_t eqt, int x) {
    /*  intercept  +  slope * x    */
    return  eqt.a  +  eqt.b * x;
}

/* let K = n
 * for each (x, y) of K:
 *      x^2, xy
 *  calc slope:
 *      m =  (K * SUM(xy) - SUM(x)*SUM(y))
 *         / (K * SUM(x^2) - SUM(x)^2)
 *  calc intercept:
 *      b = (SUM(y) - m*SUM(x))/K
 *
 * passes as pointer, NOT COPY
 */
slr_equation_t slr_find_line(int n, int X[], double Y[]) {
    //assert() len(X) == len(Y) == n
    slr_equation_t eqt;
    double sigma_x  = 0;
    double sigma_xx = 0;
    double sigma_y  = 0;
    double sigma_xy = 0;
    for (int i=0; i<n; i++) {
        sigma_x  += X[i];
        sigma_xx += pow(X[i], 2);
        sigma_y  += Y[i];
        sigma_xy += X[i] * Y[i];
    }
    double slope = regression_coefficient(n, sigma_x, sigma_xx,
                                             sigma_y, sigma_xy);
    eqt.a = y_intercept(n, slope, sigma_x, sigma_y);
    eqt.b = slope;
    return eqt;
}

/* Slope aka ”Regression Coefficient”
 *
 * Biostatistical Analysis 5ed, JH Zar, Pages 330-337
 *  Yi = a + BXi for best fit using least squares linear regression.
 *
 */
double regression_coefficient(
        int    data_points,
        double sum_of_x,        // SUM x
        double sum_sqr_of_x,    // SUM xx
        double sum_of_y,        // SUM y
        double sum_of_xy)       // SUM xy
{
    double sum_of_cross_products = sum_of_xy - sum_of_x * sum_of_y/data_points;
    double sum_of_squares_x = sum_sqr_of_x - pow(sum_sqr_of_x, 2)/data_points;
    return sum_of_cross_products / sum_of_squares_x;
}

double y_intercept(
        int    data_points,
        double slope,
        double sum_of_x,
        double sum_of_y)
{
    return (sum_of_y - slope * sum_of_x) / data_points;
}
```

Part 2: The Efficient Processor Scaling of MPI_Bcast Over Send/Recv

Program Outputs and Results

compare_bcast.c

```c
//
// Comparison of MPI_Bcast with the my_bcast function
//
#include <stdio.h>
#include <stdlib.h>
#include <assert.h>

#include <mpi.h>

/∗ send/receive process designators, maps task label to expected rank ∗/
enum TaskRanks {
    Master = 0,
    Worker = 1
};

void my_bcast(void∗ data, int count, MPI_Datatype datatype, int root,
              MPI_Comm communicator)
{
    int world_rank;
    MPI_Comm_rank(communicator, &world_rank);
    int world_size;
    MPI_Comm_size(communicator, &world_size);

    if (world_rank == root) {
        // If we are the root process, send our data to everyone


    } else {
        // If we are a receiver process, receive the data from the root


    }
}

int main(int argc, char∗∗ argv) {
    if (argc != 3) {
        fprintf(stderr, ”Usage: compare_bcast num_elements num_trials\n”);
        exit(1);
    }

    int num_elements = atoi(argv[1]);
    int num_trials = atoi(argv[2]);

    MPI_Init(NULL, NULL);

    int world_rank;
    MPI_Comm_rank(MPI_COMM_WORLD, &world_rank);

    double total_my_bcast_time = 0.0;
    double total_mpi_bcast_time = 0.0;
    int i;
    int∗ data = (int∗)malloc(sizeof(int) ∗ num_elements);
    assert(data != NULL);

    for (i = 0; i < num_trials; i++) {
        // Time my_bcast
        // Synchronize before starting timing



        // Synchronize again before obtaining final time



        // Time MPI_Bcast



    }

    // Print off timing information
    if (world_rank == 0) {
        printf(”Data size = %d, Trials = %d\n”, num_elements ∗ (int)sizeof(int),
            num_trials);
        printf(”Avg my_bcast time = %lf\n”, total_my_bcast_time / num_trials);
        printf(”Avg MPI_Bcast time = %lf\n”, total_mpi_bcast_time / num_trials);
    }

    free(data);
    MPI_Finalize();
    return 0;
}
```

hiding_latency.c

```
/* hiding_latency.c
 * ----------------
 * Authors: Darwin Jacob Groskleg
 * Date:    Saturday, May 16, 2020
 *
 * Taken from "Intro to Parallel Computing (2018), Roman Trobec et al.",
 *  Page 120
 *
 * Overlapping communication and computation
 */
#include <mpi.h>
#include <stdlib.h>
#include <math.h>
#include <stdio.h>
int i; double a;
for (i = 0; i < 100000000/numproc; i++) {
a = sin(sqrt(i)); //different amount of calculation return a;
main(int argc, char* argv[]) //number of processes must be > 1
int p, i, myid, tag=1, proc, ierr;
double start_p, run_time, start_c, comm_t, start_w, work_t, work_r; double *buff = nullptr;
MPI_Request request; MPI_Status status;
MPI_Init(&argc, &argv);
start_p = MPI_Wtime(); MPI_Comm_rank(MPI_COMM_WORLD, &myid); MPI_Comm_size(MPI_COMM_WORLD, &p);
#define master 0
#define MSGSIZE 100000000 //5000000 //different sizes of ←
messages
buff = (double*)malloc(MSGSIZE * sizeof(double)); //allocate
if (myid == master) {
for (i = 0; i < MSGSIZE; i++) { //initialize message
buff[i] = 1;
}
start_c = MPI_Wtime();
for (proc = 1; proc<p; proc++) {
MPI_Irecv(buff, MSGSIZE,
MPI_DOUBLE, MPI_ANY_SOURCE, tag, MPI_COMM_WORLD, &←
      1
2
3
4 5{ 6
   double other_work(int numproc)
   7
8 9}
   10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
29 30 31 32 33 34 35 36 37 38 39
40 41 42 43
44 45 46 47 48 49 50 51
}
int {
                            #if 1
 //non-blocking receive
  request); #endif
  #if 0
); #endif
}
```

```
MPI_Recv(buff, MSGSIZE, //blocking receive
MPI_DOUBLE, MPI_ANY_SOURCE, tag, MPI_COMM_WORLD, &status←
     }
comm_t = MPI_Wtime() - start_c; start_w = MPI_Wtime();
work_r = other_work(p);
work_t = MPI_Wtime() - start_w; MPI_Wait(&request, &status);
//block until Irecv is done

```