CS4230 Parallel Programming

Lecture 13:
Introduction to Message Passing

Mary Hall
October 23, 2012

Administrative

• Preview of next programming assignment
  - due 11:59PM, Friday, November 2
  - SVD contains several reductions
  - We will strip out the Jacobi rotations. Your mission is to implement just the reductions in MPI, using point-to-point communication and then collective communication

• Subsequent assignment
  - Scalable MPI implementation of SVD, due Friday, Nov. 9

Today's Lecture

• Message Passing, largely for distributed memory
• Message Passing Interface (MPI):
  - The most commonly-used distributed-memory programming language for large-scale computation
• Chapter 3 in textbook
• Sources for this lecture
  - Textbook slides
  - Online MPI tutorial [http://www-unix.mcs.anl.gov/mpi/tutorial/gropp/talk.html]

Recall from L3: Two main classes of parallel architecture organizations

• Shared memory multiprocessor architectures
  - A collection of autonomous processors connected to a memory system.
  - Supports a global address space where each processor can access each memory location.

• Distributed memory architectures
  - A collection of autonomous systems connected by an interconnect.
  - Each system has its own distinct address space, and processors must explicitly communicate to share data.
  - Clusters of PCs connected by commodity interconnect is the most common example.
Message Passing and MPI
• Message passing is the predominant programming model for supercomputers and clusters
  - Portable
  - Low-level, but universal and matches earlier hardware execution model
• What it is
  - A library used within conventional sequential languages (Fortran, C, C++)
  - Based on Single Program, Multiple Data (SPMD)
  - Isolation of separate address spaces
    - no data races, but communication errors possible
    - exposes execution model and forces programmer to think about locality, both good for performance
    - Complexity and code growth!
Like OpenMP, MPI arose as a standard to replace a large number of proprietary message passing libraries.

Message Passing Library Features
• All communication, synchronization require subroutine calls
  - No shared variables
  - Program runs on a single processor just like any uniprocessor program, except for calls to message passing library
• Subroutines for
  - Communication
    - Pairwise or point-to-point: A message is sent from a specific sending process (point a) to a specific receiving process (point b).
    - Collectives involving multiple processors
      - Move data: Broadcast, Scatter/gather
      - Compute and move: Reduce, AllReduce
  - Synchronization
    - Barrier
    - No locks because there are no shared variables to protect
  - Queries
    - How many processes? Which one am I? Any messages waiting?

MPI References
• The Standard itself:
  - at http://www.mpi-forum.org
  - All MPI official releases, in both postscript and HTML
• Other information on Web:
  - at http://www.mcs.anl.gov/mpi
  - pointers to lots of stuff, including other talks and tutorials, a FAQ, other MPI pages

Finding Out About the Environment
• Two important questions that arise early in a parallel program are:
  - How many processes are participating in this computation?
  - Which one am I?
• MPI provides functions to answer these questions:
  - MPI_Comm_size reports the number of processes.
  - MPI_Comm_rank reports the rank, a number between 0 and size-1, identifying the calling process
Hello (C)

```c
#include "mpi.h"
#include <stdio.h>

int main( int argc, char *argv[] )
{
    int rank, size;
    MPI_Init( &argc, &argv );
    MPI_Comm_rank( MPI_COMM_WORLD, &rank );
    MPI_Comm_size( MPI_COMM_WORLD, &size );
    printf( "Greetings from process %d of %d\n", rank, size );
    MPI_Finalize();
    return 0;
}
```

Hello (C++)

```c++
#include "mpi.h"
#include <iostream>

int main( int argc, char *argv[] )
{
    int rank, size;
    MPI::Init(argc, argv);
    rank = MPI::COMM_WORLD.Get_rank();
    size = MPI::COMM_WORLD.Get_size();
    std::cout << "Greetings from process " << rank << " of " << size << "\n";
    MPI::Finalize();
    return 0;
}
```

Compilation

```
mpicc -g -Wall -o mpi_hello mpi_hello.c
```

Create this executable file name (as opposed to default a.out)

Execution

```
mpiexec -n <number of processes> <executable>
```

```
mpiexec -n 1 /mpi_hello
```

Run with 1 process

```
mpiexec -n 4 /mpi_hello
```

Run with 4 processes
**Execution**

```
mpiexec -n 1 ./mpi_hello
Greetings from process 0 of 1!

mpiexec -n 4 ./mpi_hello
Greetings from process 0 of 4!
Greetings from process 1 of 4!
Greetings from process 2 of 4!
Greetings from process 3 of 4!
```

**MPI Components**

- **MPI_Init**
  - Tells MPI to do all the necessary setup.

  ```c
  int MPI_Init(
    int* argc, char*** argv);
  ```

- **MPI_Finalize**
  - Tells MPI we're done, so clean up anything allocated for this program.

  ```c
  int MPI_Finalize(void);
  ```

**Basic Outline**

```c
#include <mpi.h>
int main(int argc, char* argv[]) {
  /* No MPI calls before this */
  MPI_Init(&argc, &argv);
  /* MPI calls after this */
  MPI_Finalize();
  return 0;
}
```

**MPI Basic Send/Receive**

- We need to fill in the details in

  ```
  Process 0
  Send(data)  Process 1
  Receive(data)
  ```

- Things that need specifying:
  - How will "data" be described?
  - How will processes be identified?
  - How will the receiver recognize/screen messages?
  - What will it mean for these operations to complete?
MPI Basic (Blocking) Send

MPISEND(start, count, datatype, dest, tag, comm)

- The message buffer is described by (start, count, datatype).
- The target process is specified by dest, which is the rank of the target process in the communicator specified by comm.
- When this function returns, the data has been delivered to the system and the buffer can be reused. The message may not have been received by the target process.

MPI Basic (Blocking) Receive

MPI_RECV(start, count, datatype, source, tag, comm, status)

- Waits until a matching (both source and tag) message is received from the system, and the buffer can be used.
- source is rank in communicator specified by comm, or MPI_ANY_SOURCE.
- tag is a tag to be matched on or MPI_ANY_TAG.
- receiving fewer than count occurrences of datatype is OK, but receiving more is an error.
- status contains further information (e.g., size of message).

Some Basic Clarifying Concepts

- How to organize processes
  - Processes can be collected into groups.
  - Each message is sent in a context, and must be received in the same context.
  - Provides necessary support for libraries.
  - A group and context together form a communicator.
  - A process is identified by its rank in the group associated with a communicator.

- There is a default communicator whose group contains all initial processes, called MPI_COMM_WORLD.

MPI Datatypes

- The data in a message to send or receive is described by a triple (address, count, datatype), where

- An MPI datatype is recursively defined as:
  - predefined, corresponding to a data type from the language (e.g., MPI_INT, MPI_DOUBLE).
  - a contiguous array of MPI datatypes
  - a strided block of datatypes
  - an indexed array of blocks of datatypes
  - an arbitrary structure of datatypes.

- There are MPI functions to construct custom datatypes, in particular ones for subarrays.
**MPI Tags**

- Messages are sent with an accompanying user-defined integer tag, to assist the receiving process in identifying the message.
- Messages can be screened at the receiving end by specifying a specific tag, or not screened by specifying MPI_ANY_TAG as the tag in a receive.
- Some non-MPI message-passing systems have called tags "message types". MPI calls them tags to avoid confusion with datatypes.

---

**A Simple MPI Program**

```c
#include "mpi.h"
#include <stdio.h>

int main( int argc, char *argv[] )
{
    int rank, buf;
    MPI_Status status;
    MPI_Init(&argv, &argc);
    MPI_Comm_rank( MPI_COMM_WORLD, &rank );
    // Process 0 sends and Process 1 receives */
    if (rank == 0) {
        buf = 123456;
        MPI_Send( &buf, 1, MPI_INT, 1, 0, MPI_COMM_WORLD);
    } else if (rank == 1) {
        MPI_Recv( &buf, 1, MPI_INT, 0, 0, MPI_COMM_WORLD, &status );
        printf( "Received %d\n", buf );
    }
    MPI_Finalize();
    return 0;
}
```

---

**Trapezoidal Rule: Serial algorithm**

```c
/* Input:  a, b, n */

h = (b-a)/n;

approx = (f(a) + f(b))/2.0;
for (i = 1; i <= n-1; i++) {
    x_i = a + i*h;
    approx += f(x_i);
}

approx = h*approx;
```

---

**Parallel pseudo-code (naive)**

```c
Get a, b, n;

h = (b-a)/n;
local_n = n/comm_sz;
local_a = a + my_rank*local_n;
local_b = local_a + local_n;
local_integral = Trap(local_a, local_b, local_n, h);
if (my_rank == 0) {
    Send local_integral to process 0;
} else if (my_rank == 0) {
    for (proc = 1; proc < comm_sz; proc++) {
        Receive local_integral from proc;
        total_integral += local_integral;
    }
}
if (my_rank == 0) {
    print result;
}
```
First version (1)

```c
int main(void) {
    int my_rank, comm_sz, n = 1024, local_a;
    double s = 0.0, b = 3.0, h, local_s, local_b;
    double local_int, total_int;
    int source;
    MPI_Init(NULL, NULL);
    MPI_Comm_rank(MPI_COMM_WORLD, &my_rank);
    MPI_Comm_size(MPI_COMM_WORLD, &comm_sz);
    b = (b-a)/h; /* b is the same for all processes */
    local_a = a/comm_sz; /* So is the number of trapezoids */
    local_b = local_a + local_a*h;
    local_int = Trapezoid(a, b, local_s, local_b, local_a, h);
    h = (local_int - source) / (total_int - source);
    if (my_rank == 0) {
        printf("%d trapezoids, our estimate %.2f, exact is %.2f\n", n, source, exact);
        exit(0);
    }
    MPI_Finalize();
    return 0;
}
```

Copyright © 2010, Elsevier Inc. All rights Reserved

First version (2)

```c
int main(void) {
    int my_rank, comm_sz, n = 1024, local_a;
    double s = 0.0, b = 3.0, h, local_s, local_b;
    double local_int, total_int;
    int source;
    MPI_Init(NULL, NULL);
    MPI_Comm_rank(MPI_COMM_WORLD, &my_rank);
    MPI_Comm_size(MPI_COMM_WORLD, &comm_sz);
    b = (b-a)/h; /* b is the same for all processes */
    local_a = a/comm_sz; /* So is the number of trapezoids */
    local_b = local_a + local_a*h;
    local_int = Trapezoid(a, b, local_s, local_b, local_a, h);
    h = (local_int - source) / (total_int - source);
    if (my_rank == 0) {
        printf("%d trapezoids, our estimate %.2f, exact is %.2f\n", n, source, exact);
        exit(0);
    }
    MPI_Finalize();
    return 0;
}
```

Copyright © 2010, Elsevier Inc. All rights Reserved

First version (3)

```c
double Trap();
double left_endpt, right_endpt; /* in */
int trap_count; /* in */
double base_len; /* in */
double estimate, x;
int i;

estimate = (right_endpt - left_endpt)/2.0;
for (i = 1; i <= trap_count-1; i++) {
    x = left_endpt + i*base_len;
    estimate += f(x);
}
return estimate;
```

Copyright © 2010, Elsevier Inc. All rights Reserved

MPI Reduce

```c
int MPI_Reduce()
void input_data_p; /* in */
void output_data_p; /* out */
int count; /* in */
MPI_Datatype datatype; /* in */
MPI_Op operator; /* in */
int dest_process; /* in */
MPI_Comm comm; /* in */

MPI_Reduce(&local_int, &total_int, 1, MPI_DOUBLE, MPI_SUM, 0, MPI_COMM_WORLD);
```

```c
double local_a[3], b[3], sum[3];
MPI_Reduce(local_a, sum, 3, MPI_DOUBLE, MPI_SUM, 0, MPI_COMM_WORLD);
```
Replace with reduction: OpenMP version

\[
\begin{align*}
  h &= \frac{(a+b) + f(a)}{2.0} \\
  \text{approx} &= \frac{(a+b) + f(a)}{2.0} \\
  \text{for } i = 1; i < n; i++ \\
  \text{approx} &= \text{approx} + \frac{(a+b) + f(a)}{2.0} \\
  h &= \frac{(a+b) + f(a)}{2.0} \\
  \text{approx} &= \frac{(a+b) + f(a)}{2.0} \\
  \# \text{pragma omp parallel for num_threads(thread.count)} \{ \\
  \text{reduction(+): approx} \\
  \text{for } i = 1; i < n; i++ \\
  \text{approx} &= \text{approx} + \frac{(a+b) + f(a)}{2.0} \\
  \text{approx} &= \text{approx}
\end{align*}
\]

Predefined reduction operators in MPI

<table>
<thead>
<tr>
<th>Operation Value</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPI_MAX</td>
<td>Maximum</td>
</tr>
<tr>
<td>MPI_MIN</td>
<td>Minimum</td>
</tr>
<tr>
<td>MPI_SUM</td>
<td>Sun</td>
</tr>
<tr>
<td>MPI_PROD</td>
<td>Product</td>
</tr>
<tr>
<td>MPI_LAND</td>
<td>Logical and</td>
</tr>
<tr>
<td>MPI_BAND</td>
<td>Bitwise and</td>
</tr>
<tr>
<td>MPI_LOR</td>
<td>Logical or</td>
</tr>
<tr>
<td>MPI_BOR</td>
<td>Bitwise or</td>
</tr>
<tr>
<td>MPI_LOGOR</td>
<td>Logical exclusive or</td>
</tr>
<tr>
<td>MPI_EXOR</td>
<td>Bitwise exclusive or</td>
</tr>
<tr>
<td>MPI_MAXLOC</td>
<td>Maximum and location of maximum</td>
</tr>
<tr>
<td>MPI_MINLOC</td>
<td>Minimum and location of minimum</td>
</tr>
</tbody>
</table>

MPI also has reduction

```c
int MPI_Reduce(void* sendbuf, void* recvbuf, int count, MPI_Datatype datatype, MPI_Op op, int root, MPI_Comm comm)
```

```c
MPI_REDUCE(sendbuf, recvbuf, count, datatype, op, root, comm)
```

Collective vs. Point-to-Point Communications

- All the processes in the communicator must call the same collective function.
- For example, a program that attempts to match a call to `MPI_Reduce` on one process with a call to `MPI_Recv` on another process is erroneous, and, in all likelihood, the program will hang or crash.
Collective vs. Point-to-Point Communications

- The arguments passed by each process to an MPI collective communication must be "compatible."

- For example, if one process passes in 0 as the dest_process and another passes in 1, then the outcome of a call to MPI_Reduce is erroneous, and, once again, the program is likely to hang or crash.

Collective vs. Point-to-Point Communications

- The output_data_p argument is only used on dest_process.

- However, all of the processes still need to pass in an actual argument corresponding to output_data_p, even if it's just NULL.

Collective vs. Point-to-Point Communications

- Point-to-point communications are matched on the basis of tags and communicators.

- Collective communications don't use tags.

- They're matched solely on the basis of the communicator and the order in which they're called.

Next Time

- More detail on communication constructs
  - Blocking vs. non-blocking
  - One-sided communication
- Support for data and task parallelism