Announcements

• No Homework over Break
• Did post handout for the project
HW#6 Review

- Coarse

<table>
<thead>
<tr>
<th></th>
<th>real</th>
<th>load</th>
<th>convolve</th>
<th>combine</th>
<th>store</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0m2.013s</td>
<td>173,486</td>
<td>1,535,712</td>
<td>155,207</td>
<td>68,417</td>
</tr>
<tr>
<td>2</td>
<td>0m1.321s</td>
<td>144,335</td>
<td>854,183</td>
<td>200,480</td>
<td>69,430</td>
</tr>
</tbody>
</table>

- Fine
<table>
<thead>
<tr>
<th>fine:</th>
<th>real</th>
<th>load</th>
<th>conv</th>
<th>comb</th>
<th>store</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0m1.899s</td>
<td>144,764</td>
<td>1,560,142</td>
<td>109,010</td>
<td>39,953</td>
</tr>
<tr>
<td>2</td>
<td>0m1.067s</td>
<td>144,665</td>
<td>794,610</td>
<td>56,179</td>
<td>39,717</td>
</tr>
<tr>
<td>4</td>
<td>0m0.675s</td>
<td>146,151</td>
<td>364,167</td>
<td>40,649</td>
<td>53,259</td>
</tr>
<tr>
<td>8</td>
<td>0m0.486s</td>
<td>145,470</td>
<td>228,161</td>
<td>18,255</td>
<td>43,320</td>
</tr>
<tr>
<td>16</td>
<td>0m0.470s</td>
<td>146,099</td>
<td>201,600</td>
<td>13,735</td>
<td>54,268</td>
</tr>
<tr>
<td>32</td>
<td>0m0.530s</td>
<td>144,969</td>
<td>284,562</td>
<td>8,026</td>
<td>42,780</td>
</tr>
<tr>
<td>64</td>
<td>0m8.107s</td>
<td>144,948</td>
<td>7,833,613</td>
<td>21,380</td>
<td>55,849</td>
</tr>
</tbody>
</table>

- Static/Dynamic with 16 threads

<table>
<thead>
<tr>
<th></th>
<th>real</th>
<th>load</th>
<th>conv</th>
<th>comb</th>
<th>store</th>
</tr>
</thead>
<tbody>
<tr>
<td>static</td>
<td>0m0.470s</td>
<td>146,099</td>
<td>201,600</td>
<td>13,735</td>
<td>54,268</td>
</tr>
<tr>
<td>dynamic</td>
<td>0m3.776s</td>
<td>146,104</td>
<td>2,318,621</td>
<td>1,189,034</td>
<td>70,071</td>
</tr>
</tbody>
</table>
Shared Memory vs Distributed Systems

Reminder: shared memory has one copy of OS and all programs see one unified memory space.
Shared Memory

- OpenMP is nice to use. But what if your problem won’t fit on a single machine?

- How big can a shared-memory machine be?

- SGI UV systems at least 4096 cores and 16TB running one Linux image

• Digression about SGI

• Use special NUMA-Linux architecture to spread cache coherence across multiple machines.

• Origin TM and Onyx2 TM Theory of Operations Manual
Linux limitations

- Linux currently maxes out to 4096 or so.

- Somewhat dated “Scaling Linux to the Extreme” paper problems: cache contention could bring machine to halt (if a global idle counter, each thread trying to increment once a second)
  lock contention, page cache

- What are the challenges? Locking contention?

- Benefits?
(Relatively) easy to code?
Easier to port code
Many libraries do it for you. For example, OpenBLAS.
Eventually you hit the limit

What’s the alternative?
Distributed System

- Communicate over a network
- Many systems each with own memory, communicate via Message passing
- Each node has own copy of operating system
- How do they communicate?
Network Topology

- Packet-switching vs bus
- Ring, mesh, star, line, tree, fully connected
- Cube, hypercube
- Mesh networks and routing
- Routing. Fully connected? Crossbar?
Network Types

- Latency vs Bandwidth
- Top500 in Jun 2015:

<table>
<thead>
<tr>
<th>interconnect</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>infiniband FDR</td>
<td>160</td>
</tr>
<tr>
<td>10GB ethernet</td>
<td>83</td>
</tr>
<tr>
<td>infiniband QDR</td>
<td>73</td>
</tr>
<tr>
<td>gigabit ethernet</td>
<td>63</td>
</tr>
<tr>
<td>Cray Gemini</td>
<td>15</td>
</tr>
</tbody>
</table>

- Ethernet – 10/100/1Gb/10GB/40Gb/s
- **InfiniBand** – low latency, most common in supercomputers
  - copper or fiber, GB/s

<table>
<thead>
<tr>
<th></th>
<th>SDR</th>
<th>DDR</th>
<th>QDR</th>
<th>FDR-10</th>
<th>FDR</th>
<th>EDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>4x link</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>40</td>
<td>54</td>
<td>96</td>
</tr>
<tr>
<td>12x link</td>
<td>23</td>
<td>48</td>
<td>96</td>
<td>120</td>
<td>163</td>
<td>290</td>
</tr>
</tbody>
</table>

- **Cray Gemini** – Mesh/torus – 64Gb/s
- **Fibrechannel**
- **Older:** custom, Myrinet
Programming a distributed System

- Can you implement by hand?
- Sort of how you can use pthread directly?
- Yes, use ssh (like rsh) to run copy of your program on all machines
- Then write custom network code to open sockets and communicate among them all
- Network code is a pain
- Just crying out for abstraction
Message Passing Interface (MPI)

Abstraction for sending chunks of data around network. You can put together an array of 100 floats, and say "send this to process Y" and like magic it appears there.
MPI

- Message Passing Interface
- Distributed Systems
- MPI 1.0 – 1994. MPI 3.0 – 2012
- MPI 1.2 widely used. MPI 2.0 is complicated and adoption not as high as it could be.
- MPICH – CH stands for Chameleon – Argonne and Mississippi State
• MVAPICH – from Ohio State, based on MPICH

• OpenMPI – merger of 3 MPI implementations: FT-MPI from the University of Tennessee, LA-MPI from Los Alamos National Laboratory, and LAM/MPI from Indiana University

• Any other options? PVM was a predecessor

• Python Bindings, Java bindings, Matlab
MPI

Some references

https://computing.llnl.gov/tutorials/mpi/

Writing MPI code

• `#include "mpi.h"`
• Over 430 routines
• use `mpicc` to compile
  gcc or other compiler underneath, just sets up includes and libraries for you.
• `mpirun -n 4 ./test_mpi`
• `MPI_Init()` called before anything else
• `MPI_Finalize()` at the end
• Error handling – most errors just abort
Communicators

- You can specify communicator groups, and only send messages to specific groups.

- MPI_COMM_WORLD is the default, means all processes.
Rank

- Rank is the process number.
- `MPI_Comm_rank(MPI_Comm comm, int size)`
  
  ```c
  MPI_Comm_rank(MPI_COMM_WORLD, &rank);
  ```
- You can find the number of processes:
  
  ```c
  MPI_Comm_size(MPI_Comm comm, int size)
  ```
Error Handling

- MPI_SUCCESS (0) is good
- By default it aborts if any sort of error
- Can override this
Timing

- MPI_Wtime(); wallclock time in double floating point. For PAPI-like measurements
-(MPI_Wtick());
Point to Point Operations

- Buffering – what happens if we do a send but receiving side not ready?

- Blocking – blocking calls returns after it is safe to modify your send buffer. Not necessarily mean it has been sent, may just have been buffered to send. Blocking receive means only returns when all data received.

- Non-blocking – return immediately. Not safe to change buffers until you know it is finished. Wait routines for
this.

• Order – messages will not overtake each other. Send #1 and #2 to same receive, #1 will be received first

• Fairness – no guarantee of fairness. Process 1 and 2 both send to same receive on 3. No guarantee which one is received
MPI_Send, MPI_Recv

- block – MPI_Send(buffer, count, type, dest, tag, comm)
- non-block – MPI_Isend(buffer, count, type, dest, tag, comm, request)
- block – MPI_Recv(buffer, count, type, source, tag, comm, status)
- non-block – MPI_Irecv(buffer, count, type, source, tag, comm, request)

- buffer – pointer to the data buffer
- count – number of items to send
• type – MPI predefines a bunch. MPI_CHAR, MPI_INT, MPI_LONG, MPI_DOUBLE, etc. can also create own complex data types

• destination – rank to send it to

• source – rank to receive from. Also can be MPI_ANY_SOURCE

• Tag – arbitrary integer uniquely identifying message. Can pick yourself. 0-32767 guaranteed, can be higher.

• Communicator – can specify subgroups. Usually use
MPI_COMM_WORLD

- status – status of message, a struct in C
- request – on non-blocking this is a handle to the request that can be queried later to see that status
Fancier blocking send/receives

- Lots, with various type of blocking and buffer attaching and synchronous/asynchronous
/* MPI Send Example */
#include <stdio.h>
#include "mpi.h"

#define ARAYSIZE 1024*1024

int main(int argc, char **argv) {

    int numtasks, rank;
    int result,i;
    int A[ARRAYSIZE];
    MPI_Status Stat;
    int count;

    result = MPI_Init(&argc,&argv);
    if (result != MPI_SUCCESS) {
        printf ("Error starting MPI program!.
"");
        MPI_Abort(MPI_COMM_WORLD, result);
    }

    MPI_Comm_size(MPI_COMM_WORLD,&numtasks);

    // More code goes here...
}
MPI_Comm_rank(MPI_COMM_WORLD,&rank);

printf("Number of tasks= %d My rank= %d\n", numtasks, rank);

if (rank==0) {
   /* Initialize Array */
   printf("Initializing array\n");
   for(i=0;i<ARRAYSIZE;i++) {
      A[i]=1;
   }

   for(i=1;i<numtasks;i++) {
      printf("Sending %d ints to %d\n", ARRAYSIZE,i);
      result = MPI_Send(A, /* buffer */ ARRAYSIZE, /* count */
                        MPI_INT, /* type */
                        i,       /* destination */
                        13,     /* tag */
                        MPI_COMM_WORLD);
   }
}
else {
}
result = MPI_Recv(A, /* buffer */
    ARRAYSIZE, /* count */
    MPI_INT, /* type */
    0, /* source */
    13, /* tag */
    MPI_COMM_WORLD,
    &Stat);
result = MPI_Get_count(&Stat, MPI_INT, &count);
printf("\tTask %d: Received %d ints from task %d with tag %d \n",
    rank, count, Stat.MPI_SOURCE, Stat.MPI_TAG);
}

int sum=0,remote_sum=0;

for(i=rank*(ARRAYSIZE/numtasks);i<(rank+1)*(ARRAYSIZE/numtasks);i++) {
    sum+=A[i];
}

if (rank==0) {

    for(i=1;i<numtasks;i++) {
        result = MPI_Recv(&remote_sum, /* buffer */
            1, /* count */
            MPI_INT, /* type */
            MPI_COMM_WORLD,
            remote_sum, /* buffer */
            1, /* count */
            MPI_INT, /* type */
            &Stat);
        result = MPI_Get_count(&Stat, MPI_INT, &count);
        printf("\tTask %d: Received %d ints from task %d with tag %d \n",
            rank, count, Stat.MPI_SOURCE, Stat.MPI_TAG);
    }
}
MPI_ANY_SOURCE, /* source */
13, /* tag */
MPI_COMM_WORLD,
&Stat);
result = MPI_Get_count(&Stat, MPI_INT, &count);
printf("\tTask %d: (%d) Received %d int from task %d with tag %d \n",
rank,remote_sum,count,Stat.MPI_SOURCE,Stat.MPI_TAG);
sum+=remote_sum;
}

printf("Total: %d\n",sum);

}
else {
printf("\tRank %d Sending %d\n",rank,sum);
result = MPI_Send(&sum, /* buffer */
1, /* count */
MPI_INT, /* type */
0, /* destination */
13, /* tag */
MPI_COMM_WORLD);
}

MPI_Finalize();