Name: Ashwin Raghav Mohan Ganesh
UVA Id: am2qa

Assignment 1 -- CS 6444 -- Sequential Program Optimization

1.0 Introduction:

The challenge is to optimize a sequential program and gather run-time stats from the PBS cluster. Also, it is noted that the opimization is to be made to the sequential program without refactoring and making a threaded version. The objective of the exercise is to understand the operations of the GCC compiler and how these translate into the performance of an application. Another objective is to use the PBS cluster and understand the operations that can be performed on it.

2.0 Hardware and Software

Development Machines (2 configs tested):

2.1)
Dedicated personal Laptop
OS : Mac OSX Lion
Main Memory: 3GB DDR RAM
CPU: Intel x86-64bit I5

2.2)
Shared Department Systems (power1..6)
CS Department systems
OS :Ubuntu 10.04.3
Main Memory: 2GB DDR RAM
CPU: Intel x86-64bit XEON Quad Core

2.3)
Centurion001
CPU: AMD Opteron(tm) Processor 242
Main Memory 2GB RAM per node
OS :Ubuntu 10.04.3

Note: Disk/IO speed is inconsequential since the program has no CRUD operations.

3.0 Optimizations that worked well:

The code commit logs of these optimizations can be viewed incrementally in the online repository
https://github.com/ashwinraghav/Assignments/commits/master

Some of these optimizations were mentioned in the lecture and some others were the conclusions drawn from stats as shown by grpof.

3.01)

The exponentiation of the for e^a * e^b was clubbed as e^(a+b). This was the first step and I was immediately able to see a performance gain of 20%. However it is to be noted that this resulted in the precision of the resulting value being changed in the 7th decimal place (361030028.7871981263 to 361030028.7871980071). A reasonable trade-off for out particular case.

3.02)

Removed the inner loop division by ‘a’ and replace it with multiplying a constant ‘1/a’ value computed outside the loop. This converted n division operations to one division operation and n multiplication operations.

3.03)

Removed a branch in the second loop by running it till counter is ‘i’ instead of natom. This not only algorithmically reduces the time complexity (not asymptotically though) but also removes the inner condition that is crucial to let the compiler vectorize loops.

3.04)

Moved pow(cut,2) outside the loop. Brings the number of pow operations from n to 1.

3.05)

Some (3 coords[i] values and one q[i] value) array lookups were being performed on the inner loop for index values that were *invariant* inside the loop. These were moved outside to reduce the number of array lookups from 4*i to 4.

3.1 Some significant optimization lessons:

3.06)

One of the most optimizing moves was to split the second loop into 2 separate loops. One to computer Vector values and the other to conditionally increment total_e. This meant that one loop containing the 3 pow operations was not vectorizable. This improved performance by 20%.

3.07)

Another optimization step was converting the 2-dimensional double type array into a one-dimensional array and manually keeping track of the array indices. The overhead of calculating the correct index clearly won over the cache misses that would be caused in a 2D array. There was a significant performance gain once again of 25%. This particularly made a difference since the original code had array lookups in a row major manner which causes a cache miss almost every time.

4.0 Failed Optimizations

The inverse square root computation that was recommended in the lecture did not work too well on all machines. It was noted infact that in both development environments and on the cluster there was nearly a 30% drop in performance when the division by square root was replaced by multiplication of the reciprocal's square root

4.1 Disappointment in AMD's vector library

4.11)
Since the cluster machines were observed to be AMD 64 bit machines, I was hopeful that **AMD 's libm** for optimized math operations will help increase performance. Inspite of compiling the library from source on centurion001, performance only seemed to drop by 30% when these libraries were used. Quite ironically the resulting binaries were faster on my development Intel processors as opposed to AMD's own in Centurion001.
4.12)
Converting the 3rd loop contents into a function invoked through a pointer did not vectorize it. Intuitively it seemed as though changing the branched out code into a function would be a way to cheat the compiler into not seeing the branch and vectorizing the loop. However, GCC seemed to recognise pointer aliases and failed to vectorize the 3rd loop. A feature that actually turned out to be a bug.

5.0 Performance Statistics

In the interest of brevity I have provided stats for power1 development box (no 2 in the list) and Centurion001(no 3) and skipped measuring running time with flags on my local machine(no1).

The following flags that that i narrowed down on based on intuition and not stats.

Flag	Description
ffast-math	Turned on by default for all -O options
ftree-vectorizer-verbose=5	Displays unused variables and vectorization guidelines -- highly useful to improve performance in incremental steps.
-m64	use 64 bit registers
msse2	Since AMD's opteron and Intel's Xeon support sse.

-march=native	native ARM architecture
-mtune=native	tune the generated instructions to native ARM

5.1 Intermediate Readings Gathered on development machine

All intermediate readings were taken from power1.cs.virginia.edu
All Flags were used and optimization level was O3.

Optimization Number	**Total Execution Time (s)**	**Carried over to final set of optimizations**
without manual optimization	12.4900	Not Applicable
3.01	10.2800	yes
3.02	9.4800	yes
3.03	8.4500	yes
3.04	8.1400	yes
3.05	7.8500	yes
3.06	7.5400	yes
3.07	7.4300	yes
4.11	33.32	NO
4.12	7.51	NO

5.2 PBS config

```
#!/bin/sh
#PBS -l nodes=1:ppn=1
#PBS -l walltime=12:00:00
#PBS -o output.txt
#PBS -j oe
#PBS -m ea
#PBS -M am2qa@virginia.edu
```

5.3 Fully Optimized Readings

5.31 Centurion001--after complete manual optimization

Optimization Level	**Time to Calculate E**	**Total Execution**
3	12.3600	12.4100
2	12.3500	12.4100
1	13.2100	13.2600
no compiler optimization	20.59	20.64
no flags	21.12	21.18

5.32 power1.cs.virginia.edu

Optimization Level	**Time to calculate E**	**Total Exectution**
3	7.5400	7.5700
2	7.5800	7.6100
1	8.0100	8.0400
no flags	8.24	8.27
no compiler optimization	11.800	11.8400

6.0 Conclusion:

An important lesson I have learnt out of this exercise is that software ‘s performance must first be tuned to run well sequentially before parallelization. There are methods other than parallelization and algorithmic optimizations that can help improve the running speed of software. Also, it is important to be able to intuitively and non-heuristically be able to tell a compiler how a piece of software must be optimized on account of the number of options out there. In-depth knowledge about the target platform can help these decisions.

Another important lesson out of this exercise was that, although there are libraries out there that are vectorised/optimized, they need to be judged and used on a case by case basis. Extensive testing of the performance is a must on the target platform.

6.1 Flags used:

It is evident from the statistics that the performance drops with the level of optimization and with the usage of flags. I believe that this is indicative of the fact that the flags used were appropriate and in accordance to the specs of the target platform.

Also, it is to be noted that since no parallelization was adopted, running the software on the cluster did not result in a performace gain. In fact there was a performance drop on account of the individual CPU s being less performent than the machines in the development environment.

On my honor, I pledge that I have neither given nor received help on this assignment

```
/*Ashwin Raghav Mohan Ganesh tried optimizing this. You should give it a shot too!
*/
#include <stdio.h>
#include <stdlib.h>
#include <time.h>
#include <math.h>
#define min(a,b) ((a) > (b) ? (a) : (b)
double **alloc_2D_double(int nrows, int ncolumns);
void double_2D_array_free(double **array);

int main(int argc, char *argv[])
{
   long natom, i, j;
   long cut_count;

   /* Timer variables */
   clock_t time0, time1, time2;

   double cut, cut2;     /* Cut off for Rij in distance units */
   //double **coords;
   double *q, *vector2;
   double element0, element1, element2, element3;
   double subtotal_e, total_e, current_e, rij;
   double a, one_by_a;
   FILE *fptr;
   char *cptr;

   a = 3.2;
   one_by_a = 1/3.2;
   time0 = clock(); /*Start Time*/
   printf("Value of system clock at start = %ld\n",time0);

   /* Step 1 - obtain the filename of the coord file and the value of
    cut from the command line.
    Argument 1 should be the filename of the coord file (char).
    Argument 2 should be the cut off (float). */
   /* Quit therefore if iarg does not equal 3 = executable name,
    filename, cut off */
   if (argc != 3)
   {
      printf("ERROR: only %d command line options detected", argc-1);
      printf (" - need 2 options, filename and cutoff.\n");
      exit(1);
   }
   printf("Coordinates will be read from file: %s\n",argv[1]);

   /* Step 2 - Open the coordinate file and read the first line to
    obtain the number of atoms */
   if ((fptr=fopen(argv[1],"r"))==NULL)
   {
      printf("ERROR: Could not open file called %s\n",argv[1]);
      exit(1);
   }
   else
   {
      fscanf(fptr, "%ld", &natom);
   }

   printf("Natom = %ld\n", natom);

   cut = strtod(argv[2],&cptr);
   printf("cut = %10.4f\n", cut);

   /* Step 3 - Allocate the arrays to store the coordinate and charge
    data */
   //coords=alloc_2D_double(3,natom);
   double coords[natom*3];
   if ( coords==NULL )
   {
      printf("Allocation error coords");
      exit(1);
   }
   q=(double *)malloc(natom*sizeof(double));
   vector2=(double *)malloc(natom*sizeof(double));
   if ( q == NULL )
   {
      printf("Allocation error q");
      exit(1);
   }


   /* Step 4 - read the coordinates and charges. */
   for (i = 0; i<natom; ++i)
```

```
{
   fscanf(fptr, "%lf %lf %lf %lf",&coords[i * 3],
        &coords[(3*i)+1],&coords[(3*i)+2],&q[i]);
}

time1 = clock(); /*time after file read*/
printf("Value of system clock after coord read = %ld\n",time1);


/* Step 5 - calculate the number of pairs and E. - this is the
 majority of the work. */
total_e = 0.0;
cut_count = 0;
cut2 = pow(cut, 2);
for (i=1; i<=natom; ++i)
{
   element0 = coords[(3*(i-1))];
   element1 = coords[(3*(i-1))+1];
   element2 = coords[(3*(i-1))+2];
   element3 = q[i-1];
   for (j=1; j < i; ++j)
   {
      /* X^2 + Y^2 + Z^2 */
      vector2[j] =
      pow(element0-coords[(3*(j-1))],2.0) +
      pow(element1-coords[(3*(j-1))+1],2.0) +
      pow(element2-coords[(3*(j-1))+2],2.0);
   }
   for(j=1; j< i - 30; ++j){
      if (vector2[j] < cut2)
      {
         rij = sqrt(vector2[j]);
         ++cut_count;
         current_e = (exp(rij*(element3 + q[j-1])))/rij;
         total_e = total_e + current_e - one_by_a;
      }
   }

}

time2 = clock(); /* time after reading of file and calculation */
printf("Value of system clock after coord read and E calc = %ld\n", time2);
```

```
    /* Step 6 - write out the results */
    printf("                         Final Results\n");
    printf("                         -------------\n");
    printf("                   Num Pairs = %ld\n",cut_count);
    printf("                     Total E = %14.10f\n",total_e);
    printf("     Time to read coord file = %14.4f Seconds\n",
           ((double )(time1-time0))/(double )CLOCKS_PER_SEC);
    printf("         Time to calculate E = %14.4f Seconds\n",
           ((double )(time2-time1))/(double )CLOCKS_PER_SEC);
    printf("        Total Execution Time = %14.4f Seconds\n",
           ((double )(time2-time0))/(double )CLOCKS_PER_SEC);

    /* Step 7 - Deallocate the arrays - we should strictly check the
     return values here but for the purposes of this tutorial we can
     ignore this. */
    free(q);
    //double_2D_array_free(coords);

    fclose(fptr);

    exit(0);
}

double **alloc_2D_double(int nrows, int ncolumns)
{
    /* Allocates a 2d_double_array consisting of a series of pointers
     pointing to each row that are then allocated to be ncolumns
     long each. */

    /* Try's to keep contents contiguous - thus reallocation is
     difficult! */

    /* Returns the pointer **array. Returns NULL on error */
    int i;

    double **array = (double **)malloc(nrows*sizeof(double *));
    if (array==NULL)
        return NULL;
    array[0] = (double *)malloc(nrows*ncolumns*sizeof(double));
    if (array[0]==NULL)
        return NULL;
    for (i = 1; i < nrows; ++i)
        array[i] = array[0] + i * ncolumns;

    return array;

}

void double_2D_array_free(double **array)
{
    /* Frees the memory previously allocated by alloc_2D_double */
    free(array[0]);
    free(array);
}
// gcc -lm -O3 -g -Wall -ftree-vectorizer-verbose=5 -msse -msse2 -msse3 -march=native -mtune=native --std=c99 -fPIC -ffast-math original.c
```