Computational Physics Lectures: How to optimize codes, from vectorization to parallelization

Morten Hjorth-Jensen[1,2]

[1]Department of Physics, University of Oslo
[2]Department of Physics and Astronomy and National Superconducting Cyclotron Laboratory, Michigan State University

Oct 17, 2019

Content

- Simple compiler options
- Tools to benchmark your code
- Machine architectures
- What is vectorization?
- Parallelization with OpenMP
- Parallelization with MPI
- Vectorization and parallelization, examples

Optimization and profiling

Till now we have not paid much attention to speed and possible optimization possibilities inherent in the various compilers. We have compiled and linked as For Fortran replace with for example **gfortran** or **ifort**. This is what we call a flat compiler option and should be used when we develop the code. It produces normally a very large and slow code when translated to machine instructions. We use this option for debugging and for establishing the correct program output because every operation is done precisely as the user specified it.

It is instructive to look up the compiler manual for further instructions by writing

More on optimization

We have additional compiler options for optimization. These may include procedure inlining where performance may be improved, moving constants inside loops outside the loop, identify potential parallelism, include automatic vectorization or replace a division with a reciprocal and a multiplication if this speeds up the code. This (other options are -O2 or -Ofast) is the recommended option.

Optimization and profiling

It is also useful to profile your program under the development stage. You would then compile with After you have run the code you can obtain the profiling information via When you have profiled properly your code, you must take out this option as it slows down performance. For memory tests use valgrind. An excellent environment for all these aspects, and much more, is Qt creator.

Optimization and debugging

Adding debugging options is a very useful alternative under the development stage of a program. You would then compile with This option generates debugging information allowing you to trace for example if an array is properly allocated. Some compilers work best with the no optimization option **-O0**.

Other optimization flags. Depending on the compiler, one can add flags which generate code that catches integer overflow errors. The flag **-ftrapv** does this for the CLANG compiler on OS X operating systems.

Other hints

In general, irrespective of compiler options, it is useful to

- avoid if tests or call to functions inside loops, if possible.
- avoid multiplication with constants inside loops if possible

Here is an example of a part of a program where specific operations lead to a slower code A better code is Here we avoid a repeated multiplication inside a loop. Most compilers, depending on compiler flags, identify and optimize such bottlenecks on their own, without requiring any particular action by the programmer. However, it is always useful to single out and avoid code examples like the first one discussed here.

Vectorization and the basic idea behind parallel computing

Present CPUs are highly parallel processors with varying levels of parallelism. The typical situation can be described via the following three statements.

- Pursuit of shorter computation time and larger simulation size gives rise to parallel computing.
- Multiple processors are involved to solve a global problem.
- The essence is to divide the entire computation evenly among collaborative processors. Divide and conquer.

Before we proceed with a more detailed discussion of topics like vectorization and parallelization, we need to remind ourselves about some basic features of different hardware models.

A rough classification of hardware models

- Conventional single-processor computers are named SISD (single-instruction-single-data) machines.
- SIMD (single-instruction-multiple-data) machines incorporate the idea of parallel processing, using a large number of processing units to execute the same instruction on different data.
- Modern parallel computers are so-called MIMD (multiple-instruction-multiple-data) machines and can execute different instruction streams in parallel on different data.

Shared memory and distributed memory

One way of categorizing modern parallel computers is to look at the memory configuration.

- In shared memory systems the CPUs share the same address space. Any CPU can access any data in the global memory.
- In distributed memory systems each CPU has its own memory.

The CPUs are connected by some network and may exchange messages.

Different parallel programming paradigms

- **Task parallelism**: the work of a global problem can be divided into a number of independent tasks, which rarely need to synchronize. Monte Carlo simulations represent a typical situation. Integration is another. However this paradigm is of limited use.

- **Data parallelism**: use of multiple threads (e.g. one or more threads per processor) to dissect loops over arrays etc. Communication and synchronization between processors are often hidden, thus easy to program. However, the user surrenders much control to a specialized compiler. Examples of data parallelism are compiler-based parallelization and OpenMP directives.

Different parallel programming paradigms

- **Message passing**: all involved processors have an independent memory address space. The user is responsible for partitioning the data/work of a global problem and distributing the subproblems to the processors. Collaboration between processors is achieved by explicit message passing, which is used for data transfer plus synchronization.
- This paradigm is the most general one where the user has full control. Better parallel efficiency is usually achieved by explicit message passing. However, message-passing programming is more difficult.

What is vectorization?

Vectorization is a special case of **Single Instructions Multiple Data** (SIMD) to denote a single instruction stream capable of operating on multiple data elements in parallel. We can think of vectorization as the unrolling of loops accompanied with SIMD instructions.

Vectorization is the process of converting an algorithm that performs scalar operations (typically one operation at the time) to vector operations where a single operation can refer to many simultaneous operations. Consider the following example If the code is not vectorized, the compiler will simply start with the first element and then perform subsequent additions operating on one address in memory at the time.

Number of elements that can acted upon

A SIMD instruction can operate on multiple data elements in one single instruction. It uses the so-called 128-bit SIMD floating-point register. In this sense,vectorization adds some form of parallelism since one instruction is applied to many parts of say a vector.

The number of elements which can be operated on in parallel range from four single-precision floating point data elements in so-called Streaming SIMD Extensions and two double-precision floating-point data elements in Streaming SIMD Extensions 2 to sixteen byte operations in a 128-bit register in Streaming SIMD Extensions 2. Thus, vector-length ranges from 2 to 16, depending on the instruction extensions used and on the data type.

Number of elements that can acted upon, examples

We start with the simple scalar operations given by If the code is not vectorized and we have a 128-bit register to store a 32 bits floating point number, it means that we have 3×32 bits that are not used. For the first element we have

0	1	2	3
a[0]=	not used	not used	not used
b[0]+	not used	not used	not used
c[0]	not used	not used	not used

We have thus unused space in our SIMD registers. These registers could hold three additional integers.

Number of elements that can acted upon, examples

If we vectorize the code, we can perform, with a 128-bit register four simultaneous operations, that is we have displayed here as

0	1	2	3
a[0]=	a[1]=	a[2]=	a[3]=
b[0]+	b[1]+	b[2]+	b[3]+
c[0]	c[1]	c[2]	c[3]

Four additions are now done in a single step.

A simple test case with and without vectorization

We implement these operations in a simple c++ program as

```
#include <cstdlib>
#include <iostream>
#include <cmath>
#include <iomanip>
#include "time.h"

using namespace std; // note use of namespace
int main (int argc, char* argv[])
{
  int i = atoi(argv[1]);
  double *a, *b, *c;
  a = new double[i];
  b = new double[i];
  c = new double[i];
  for (int j = 0; j < i; j++) {
    a[j] = 0.0;
    b[j] = cos(j*1.0);
    c[j] = sin(j*3.0);
  }
  clock_t start, finish;
  start = clock();
  for (int j = 0; j < i; j++) {
    a[j] = b[j]+b[j]*c[j];
  }
  finish = clock();
```

```
  double timeused = (double) (finish - start)/(CLOCKS_PER_SEC );
  cout << setiosflags(ios::showpoint | ios::uppercase);
  cout << setprecision(10) << setw(20) << "Time used  for vector addition and multiplication=" << time
  delete [] a;
  delete [] b;
  delete [] c;
  return 0;
}
```

Compiling with and without vectorization

We can compile and link without vectorization and with vectorization (and additional optimizations) The speedup depends on the size of the vectors. In the example here we have run with 10^7 elements. The example here was run on a PC with ubuntu 14.04 as operating system and an Intel i7-4790 CPU running at 3.60 GHz. This particular C++ compiler speeds up the above loop operations with a factor of 3. Performing the same operations for 10^8 elements results only in a factor 1.4. The result will however vary from compiler to compiler. In general however, with optimization flags like $-O3$ or $-Ofast$, we gain a considerable speedup if our code can be vectorized. Many of these operations can be done automatically by your compiler. These automatic or near automatic compiler techniques improve performance considerably.

Automatic vectorization and vectorization inhibitors, criteria

Not all loops can be vectorized, as discussed in Intel's guide to vectorization

An important criteria is that the loop counter n is known at the entry of the loop. The variable n does need to be known at compile time. However, this variable must stay the same for the entire duration of the loop. It implies that an exit statement inside the loop cannot be data dependent.

Automatic vectorization and vectorization inhibitors, exit criteria

An exit statement should in general be avoided. If the exit statement contains data-dependent conditions, the loop cannot be vectorized. The following is an example of a non-vectorizable loop Avoid loop termination conditions and opt for a single entry loop variable n. The lower and upper bounds have to be kept fixed within the loop.

Automatic vectorization and vectorization inhibitors, straight-line code

SIMD instructions perform the same type of operations multiple times. A **switch** statement leads thus to a non-vectorizable loop since different statemens cannot branch. The following code can however be vectorized since the **if** statement is implemented as a masked assignment. These operations can be performed for

all data elements but only those elements which the mask evaluates as true are stored. In general, one should avoid branches such as **switch**, **go to**, or **return** statements or **if** constructs that cannot be treated as masked assignments.

Automatic vectorization and vectorization inhibitors, nested loops

Only the innermost loop of the following example is vectorized The exception is if an original outer loop is transformed into an inner loop as the result of compiler optimizations.

Automatic vectorization and vectorization inhibitors, function calls

Calls to programmer defined functions ruin vectorization. However, calls to intrinsic functions like $\sin x$, $\cos x$, $\exp x$ etc are allowed since they are normally efficiently vectorized. The following example is fully vectorizable Similarly, **inline** functions defined by the programmer, allow for vectorization since the function statements are glued into the actual place where the function is called.

Automatic vectorization and vectorization inhibitors, data dependencies

One has to keep in mind that vectorization changes the order of operations inside a loop. A so-called read-after-write statement with an explicit flow dependency cannot be vectorized. The following code is an example of flow dependency and results in wrong numerical results if vectorized. For a scalar operation, the value $a[i-1]$ computed during the iteration is loaded into the right-hand side and the results are fine. In vector mode however, with a vector length of four, the values $a[0]$, $a[1]$, $a[2]$ and $a[3]$ from the previous loop will be loaded into the right-hand side and produce wrong results. That is, we have and if the two first iterations are executed at the same by the SIMD instruction, the value of say $a[1]$ could be used by the second iteration before it has been calculated by the first iteration, leading thereby to wrong results.

Automatic vectorization and vectorization inhibitors, more data dependencies

On the other hand, a so-called write-after-read statement can be vectorized. The following code is an example of flow dependency that can be vectorized since no iteration with a higher value of i can complete before an iteration with a lower value of i. However, such code leads to problems with parallelization.

Automatic vectorization and vectorization inhibitors, memory stride

For C++ programmers it is also worth keeping in mind that an array notation is preferred to the more compact use of pointers to access array elements. The compiler can often not tell if it is safe to vectorize the code.

When dealing with arrays, you should also avoid memory stride, since this slows down considerably vectorization. When you access array element, write for example the inner loop to vectorize using unit stride, that is, access successively the next array element in memory, as shown here

Compiling with and without vectorization

We can compile and link without vectorization using the clang c++ compiler and with vectorization (and additional optimizations) The speedup depends on the size of the vectors. In the example here we have run with 10^7 elements. The example here was run on an IMac17.1 with OSX El Capitan (10.11.4) as operating system and an Intel i5 3.3 GHz CPU. This particular C++ compiler speeds up the above loop operations with a factor of 1.5 Performing the same operations for 10^9 elements results in a smaller speedup since reading from main memory is required. The non-vectorized code is seemingly faster. We will discuss these issues further in the next slides.

Compiling with and without vectorization using clang

We can compile and link without vectorization with clang compiler and with vectorization We can also add vectorization analysis, see for example or figure out if vectorization was missed

Memory management

The main memory contains the program data

- Cache memory contains a copy of the main memory data
- Cache is faster but consumes more space and power. It is normally assumed to be much faster than main memory
- Registers contain working data only
  - Modern CPUs perform most or all operations only on data in register
- Multiple Cache memories contain a copy of the main memory data
  - Cache items accessed by their address in main memory
  - L1 cache is the fastest but has the least capacity
  - L2, L3 provide intermediate performance/size tradeoffs

Loads and stores to memory can be as important as floating point operations when we measure performance.

Memory and communication

- Most communication in a computer is carried out in chunks, blocks of bytes of data that move together
- In the memory hierarchy, data moves between memory and cache, and between different levels of cache, in groups called lines
  - Lines are typically 64-128 bytes, or 8-16 double precision words
  - Even if you do not use the data, it is moved and occupies space in the cache
- This performance feature is not captured in most programming languages

Measuring performance

How do we measure erformance? What is wrong with this code to time a loop?

```
clock_t start, finish;
start = clock();
for (int j = 0; j < i; j++) {
  a[j] = b[j]+b[j]*c[j];
}
finish = clock();
double timeused = (double) (finish - start)/(CLOCKS_PER_SEC );
```

Problems with measuring time

1. Timers are not infinitely accurate
2. All clocks have a granularity, the minimum time that they can measure
3. The error in a time measurement, even if everything is perfect, may be the size of this granularity (sometimes called a clock tick)
4. Always know what your clock granularity is
5. Ensure that your measurement is for a long enough duration (say 100 times the **tick**)

Problems with cold start

What happens when the code is executed? The assumption is that the code is ready to execute. But

1. Code may still be on disk, and not even read into memory.
2. Data may be in slow memory rather than fast (which may be wrong or right for what you are measuring)
3. Multiple tests often necessary to ensure that cold start effects are not present
4. Special effort often required to ensure data in the intended part of the memory hierarchy.

Problems with smart compilers

1. If the result of the computation is not used, the compiler may eliminate the code
2. Performance will look impossibly fantastic
3. Even worse, eliminate some of the code so the performance looks plausible
4. Ensure that the results are (or may be) used.

Problems with interference

1. Other activities are sharing your processor
    - Operating system, system demons, other users
        - Some parts of the hardware do not always perform with exactly the same performance
2. Make multiple tests and report
3. Easy choices include
    - Average tests represent what users might observe over time

Problems with measuring performance

1. Accurate, reproducible performance measurement is hard
2. Think carefully about your experiment:
3. What is it, precisely, that you want to measure
4. How representative is your test to the situation that you are trying to measure?

Thomas algorithm for tridiagonal linear algebra equations

$$\begin{pmatrix} b_0 & c_0 & & & \\ a_0 & b_1 & c_1 & & \\ & & \ddots & & \\ & & a_{m-3} & b_{m-2} & c_{m-2} \\ & & & a_{m-2} & b_{m-1} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{m-2} \\ x_{m-1} \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{m-2} \\ f_{m-1} \end{pmatrix}$$

Thomas algorithm, forward substitution

The first step is to multiply the first row by a_0/b_0 and subtract it from the second row. This is known as the forward substitution step. We obtain then

$$a_i = 0,$$

$$b_i = b_i - \frac{a_{i-1}}{b_{i-1}} c_{i-1},$$

and

$$f_i = f_i - \frac{a_{i-1}}{b_{i-1}} f_{i-1}.$$

At this point the simplified equation, with only an upper triangular matrix takes the form

$$\begin{pmatrix} b_0 & c_0 & & & \\ & b_1 & c_1 & & \\ & & \ddots & & \\ & & & b_{m-2} & c_{m-2} \\ & & & & b_{m-1} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{m-2} \\ x_{m-1} \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{m-2} \\ f_{m-1} \end{pmatrix}$$

Thomas algorithm, backward substitution

The next step is the backward substitution step. The last row is multiplied by c_{N-3}/b_{N-2} and subtracted from the second to last row, thus eliminating c_{N-3} from the last row. The general backward substitution procedure is

$$c_i = 0,$$

and

$$f_{i-1} = f_{i-1} - \frac{c_{i-1}}{b_i} f_i$$

All that remains to be computed is the solution, which is the very straight forward process of

$$x_i = \frac{f_i}{b_i}$$

Thomas algorithm and counting of operations (floating point and memory)

Operation	Floating Point
Memory Reads	$14(N-2)$
Memory Writes	$4(N-2)$
Subtractions	$3(N-2)$
Multiplications	$3(N-2)$
Divisions	$4(N-2)$

An inefficient code.

The specialized Thomas algorithm (Project 1)

Operation	Floating Point
Memory Reads	$6(N-2)$
Memory Writes	$2(N-2)$
Additions	$2(N-2)$
Divisions	$2(N-2)$

Example: Transpose of a matrix

```
#include <cstdlib>
#include <iostream>
#include <cmath>
#include <iomanip>
#include "time.h"

using namespace std; // note use of namespace
int main (int argc, char* argv[])
{
  // read in dimension of square matrix
  int n = atoi(argv[1]);
  double **A, **B;
  // Allocate space for the two matrices
  A = new double*[n]; B = new double*[n];
  for (int i = 0; i < n; i++){
    A[i] = new double[n];
    B[i] = new double[n];
  }
  // Set up values for matrix A
  for (int i = 0; i < n; i++){
    for (int j = 0; j < n; j++) {
      A[i][j] =  cos(i*1.0)*sin(j*3.0);
    }
  }
  clock_t start, finish;
  start = clock();
```

```
  // Then compute the transpose
  for (int i = 0; i < n; i++){
    for (int j = 0; j < n; j++) {
      B[i][j]= A[j][i];
    }
  }

  finish = clock();
  double timeused = (double) (finish - start)/(CLOCKS_PER_SEC );
  cout << setiosflags(ios::showpoint | ios::uppercase);
  cout << setprecision(10) << setw(20) << "Time used  for setting up transpose of matrix=" << timeused

  // Free up space
  for (int i = 0; i < n; i++){
    delete[] A[i];
    delete[] B[i];
  }
  delete[] A;
  delete[] B;
  return 0;
}
```

Matrix-matrix multiplication

This the matrix-matrix multiplication code with plain c++ memory allocation.
It computes at the end the Frobenius norm.

```
#include <cstdlib>
#include <iostream>
#include <cmath>
#include <iomanip>
#include "time.h"

using namespace std; // note use of namespace
int main (int argc, char* argv[])
{
  // read in dimension of square matrix
  int n = atoi(argv[1]);
  double s = 1.0/sqrt( (double) n);
  double **A, **B, **C;
  // Start timing
  clock_t start, finish;
  start = clock();
  // Allocate space for the two matrices
  A = new double*[n]; B = new double*[n]; C = new double*[n];
  for (int i = 0; i < n; i++){
    A[i] = new double[n];
    B[i] = new double[n];
    C[i] = new double[n];
  }
  // Set up values for matrix A and B and zero matrix C
  for (int i = 0; i < n; i++){
    for (int j = 0; j < n; j++) {
      double angle = 2.0*M_PI*i*j/ (( double ) n);
      A[i][j] = s * ( sin ( angle ) + cos ( angle ) );
      B[j][i] =  A[i][j];
    }
  }
  // Then perform the matrix-matrix multiplication
  for (int i = 0; i < n; i++){
    for (int j = 0; j < n; j++) {
```

```
      double sum = 0.0;
       for (int k = 0; k < n; k++) {
           sum += B[i][k]*A[k][j];
       }
       C[i][j] = sum;
    }
  }
  // Compute now the Frobenius norm
  double Fsum = 0.0;
  for (int i = 0; i < n; i++){
    for (int j = 0; j < n; j++) {
      Fsum += C[i][j]*C[i][j];
    }
  }
  Fsum = sqrt(Fsum);
  finish = clock();
  double timeused = (double) (finish - start)/(CLOCKS_PER_SEC );
  cout << setiosflags(ios::showpoint | ios::uppercase);
  cout << setprecision(10) << setw(20) << "Time used  for matrix-matrix multiplication=" << timeused
  cout << "  Frobenius norm  = " << Fsum << endl;
  // Free up space
  for (int i = 0; i < n; i++){
    delete[] A[i];
    delete[] B[i];
    delete[] C[i];
  }
  delete[] A;
  delete[] B;
  delete[] C;
  return 0;
}
```

How do we define speedup? Simplest form

- Speedup(code,sys,p) = T_b/T_p
- Speedup measures the ratio of performance between two objects
- Versions of same code, with different number of processors
- Serial and vector versions
- Try different programing languages, c++ and Fortran
- Two algorithms computing the **same** result

How do we define speedup? Correct baseline

The key is choosing the correct baseline for comparison

- For our serial vs. vectorization examples, using compiler-provided vectorization, the baseline is simple; the same code, with vectorization turned off
  - For parallel applications, this is much harder:

* Choice of algorithm, decomposition, performance of baseline case etc.

Parallel speedup

For parallel applications, speedup is typically defined as

- Speedup(code,sys,p) = T_1/T_p

Here T_1 is the time on one processor and T_p is the time using p processors.

- Can Speedup(code,sys,p) become larger than p?

That means using p processors is more than p times faster than using one processor.

Speedup and memory

The speedup on p processors can be greater than p if memory usage is optimal! Consider the case of a memorybound computation with M words of memory

- If M/p fits into cache while M does not, the time to access memory will be different in the two cases:
- T_1 uses the main memory bandwidth
- T_p uses the appropriate cache bandwidth

Upper bounds on speedup

Assume that almost all parts of a code are perfectly parallelizable (fraction f). The remainder, fraction $(1-f)$ cannot be parallelized at all.

That is, there is work that takes time W on one process; a fraction f of that work will take time Wf/p on p processors.

- What is the maximum possible speedup as a function of f?

Amdahl's law

On one processor we have

$$T_1 = (1-f)W + fW = W$$

On p processors we have

$$T_p = (1-f)W + \frac{fW}{p},$$

resulting in a speedup of

$$\frac{T_1}{T_p} = \frac{W}{(1-f)W + fW/p}$$

As p goes to infinity, fW/p goes to zero, and the maximum speedup is

$$\frac{1}{1-f},$$

meaning that if if $f = 0.99$ (all but 1% parallelizable), the maximum speedup is $1/(1-.99) = 100$!

How much is parallelizable

If any non-parallel code slips into the application, the parallel performance is limited.

In many simulations, however, the fraction of non-parallelizable work is 10^{-6} or less due to large arrays or objects that are perfectly parallelizable.

Today's situation of parallel computing

- Distributed memory is the dominant hardware configuration. There is a large diversity in these machines, from MPP (massively parallel processing) systems to clusters of off-the-shelf PCs, which are very cost-effective.
- Message-passing is a mature programming paradigm and widely accepted. It often provides an efficient match to the hardware. It is primarily used for the distributed memory systems, but can also be used on shared memory systems.
- Modern nodes have nowadays several cores, which makes it interesting to use both shared memory (the given node) and distributed memory (several nodes with communication). This leads often to codes which use both MPI and OpenMP.

Our lectures will focus on both MPI and OpenMP.

Overhead present in parallel computing

- **Uneven load balance**: not all the processors can perform useful work at all time.
- **Overhead of synchronization**
- **Overhead of communication**

- **Extra computation due to parallelization**

Due to the above overhead and that certain parts of a sequential algorithm cannot be parallelized we may not achieve an optimal parallelization.

Parallelizing a sequential algorithm

- Identify the part(s) of a sequential algorithm that can be executed in parallel. This is the difficult part,
- Distribute the global work and data among P processors.

Strategies

- Develop codes locally, run with some few processes and test your codes. Do benchmarking, timing and so forth on local nodes, for example your laptop or PC.
- When you are convinced that your codes run correctly, you can start your production runs on available supercomputers.

How do I run MPI on a PC/Laptop? MPI

To install MPI is rather easy on hardware running unix/linux as operating systems, follow simply the instructions from the OpenMPI website. See also subsequent slides. When you have made sure you have installed MPI on your PC/laptop,

- Compile with mpicxx/mpic++ or mpif90

Can I do it on my own PC/laptop? OpenMP installation

If you wish to install MPI and OpenMP on your laptop/PC, we recommend the following:

- For OpenMP, the compile option **-fopenmp** is included automatically in recent versions of the C++ compiler and Fortran compilers. For users of different Linux distributions, siply use the available C++ or Fortran compilers and add the above compiler instructions, see also code examples below.
- For OS X users however, install **libomp**

and compile and link as

Installing MPI

For linux/ubuntu users, you need to install two packages (alternatively use the synaptic package manager) For OS X users, install brew (after having installed xcode and gcc, needed for the gfortran compiler of openmpi) and then install with brew When running an executable (code.x), run as where we indicate that we want the number of processes to be 10.

Installing MPI and using Qt

With openmpi installed, when using Qt, add to your .pro file the instructions here

You may need to tell Qt where openmpi is stored.

For the machines at the computer lab, openmpi is located at Add to your *.bashrc* file the following

Using Smaug, the CompPhys computing cluster

For running on SMAUG, go to `http://comp-phys.net/` and click on the link internals and click on computing cluster. To get access to Smaug, you will need to send us an e-mail with your name, UiO username, phone number, room number and affiliation to the research group. In return, you will receive a password you may use to access the cluster.

Here follows a simple recipe In the folder you will find a simple example on how to set up a job and compile and run. This files are write protected. Copy them to your own folder and compile and run there. For more information see the readme file under the program folder.

What is OpenMP

- OpenMP provides high-level thread programming
- Multiple cooperating threads are allowed to run simultaneously
- Threads are created and destroyed dynamically in a fork-join pattern
  - An OpenMP program consists of a number of parallel regions
  - Between two parallel regions there is only one master thread
  - In the beginning of a parallel region, a team of new threads is spawned
- The newly spawned threads work simultaneously with the master thread
- At the end of a parallel region, the new threads are destroyed

Many good tutorials online and excellent textbook

1. Using OpenMP, by B. Chapman, G. Jost, and A. van der Pas

2. Many tutorials online like OpenMP official site

Getting started, things to remember

- Remember the header file
- Insert compiler directives in C++ syntax as
- Compile with for example *c++ -fopenmp code.cpp*
- Execute
  - Remember to assign the environment variable **OMP NUM THREADS**
  - It specifies the total number of threads inside a parallel region, if not otherwise overwritten

OpenMP syntax

- Mostly directives
- Some functions and types
- Most apply to a block of code
- Specifically, a **structured block**
- Enter at top, exit at bottom only, exit(), abort() permitted

Different OpenMP styles of parallelism

OpenMP supports several different ways to specify thread parallelism

- General parallel regions: All threads execute the code, roughly as if you made a routine of that region and created a thread to run that code
- Parallel loops: Special case for loops, simplifies data parallel code
- Task parallelism, new in OpenMP 3
- Several ways to manage thread coordination, including Master regions and Locks
- Memory model for shared data

General code structure

Parallel region

- A parallel region is a block of code that is executed by a team of threads
- The following compiler directive creates a parallel region
- Clauses can be added at the end of the directive
- Most often used clauses:
  - **default(shared)** or **default(none)**
  - **public(list of variables)**
  - **private(list of variables)**

Hello world, not again, please!

Hello world, yet another variant

Variables declared outside of the parallel region are shared by all threads If a variable like **id** is declared outside of the it would have been shared by various the threads, possibly causing erroneous output

- Why? What would go wrong? Why do we add possibly?

Important OpenMP library routines

- **int omp get num threads ()**, returns the number of threads inside a parallel region
- **int omp get thread num ()**, returns the a thread for each thread inside a parallel region
- **void omp set num threads (int)**, sets the number of threads to be used
- **void omp set nested (int)**, turns nested parallelism on/off

Private variables

Private clause can be used to make thread- private versions of such variables:

- What is their value on entry? Exit?
- OpenMP provides ways to control that
- Can use default(none) to require the sharing of each variable to be described

Master region

It is often useful to have only one thread execute some of the code in a parallel region. I/O statements are a common example

Parallel for loop

- Inside a parallel region, the following compiler directive can be used to parallelize a for-loop:
- Clauses can be added, such as
  - **schedule(static, chunk size)**
  - **schedule(dynamic, chunk size)**
  - **schedule(guided, chunk size)** (non-deterministic allocation)
  - **schedule(runtime)**
  - **private(list of variables)**
  - **reduction(operator:variable)**
  - **nowait**

Parallel computations and loops

OpenMP provides an easy way to parallelize a loop OpenMP handles index variable (no need to declare in for loop or make private)

Which thread does which values? Several options.

Scheduling of loop computations

We can let the OpenMP runtime decide. The decision is about how the loop iterates are scheduled and OpenMP defines three choices of loop scheduling:

1. Static: Predefined at compile time. Lowest overhead, predictable
2. Dynamic: Selection made at runtime
3. Guided: Special case of dynamic; attempts to reduce overhead

Example code for loop scheduling

Example code for loop scheduling, guided instead of dynamic

More on Parallel for loop

- The number of loop iterations cannot be non-deterministic; break, return, exit, goto not allowed inside the for-loop
- The loop index is private to each thread
- A reduction variable is special
  - During the for-loop there is a local private copy in each thread
  - At the end of the for-loop, all the local copies are combined together by the reduction operation
- Unless the nowait clause is used, an implicit barrier synchronization will be added at the end by the compiler

can be combined into

What can happen with this loop?

What happens with code like this All threads can access the **sum** variable, but the addition is not atomic! It is important to avoid race between threads. So-called reductions in OpenMP are thus important for performance and for obtaining correct results. OpenMP lets us indicate that a variable is used for a reduction with a particular operator. The above code becomes

Inner product

$$\sum_{i=0}^{n-1} a_i b_i$$

Different threads do different tasks

Different threads do different tasks independently, each section is executed by one thread.

Single execution

The code is executed by one thread only, no guarantee which thread
Can introduce an implicit barrier at the end Code executed by the master thread, guaranteed and no implicit barrier at the end.

Coordination and synchronization

Synchronization, must be encountered by all threads in a team (or none) is another form of synchronization (in sequential order). The form and is more efficient than

Data scope

- OpenMP data scope attribute clauses:
  - **shared**
  - **private**
  - **firstprivate**
  - **lastprivate**
  - **reduction**

What are the purposes of these attributes

- define how and which variables are transferred to a parallel region (and back)
- define which variables are visible to all threads in a parallel region, and which variables are privately allocated to each thread

Some remarks

- When entering a parallel region, the **private** clause ensures each thread having its own new variable instances. The new variables are assumed to be uninitialized.
- A shared variable exists in only one memory location and all threads can read and write to that address. It is the programmer's responsibility to ensure that multiple threads properly access a shared variable.
- The **firstprivate** clause combines the behavior of the private clause with automatic initialization.
- The **lastprivate** clause combines the behavior of the private clause with a copy back (from the last loop iteration or section) to the original variable outside the parallel region.

Parallelizing nested for-loops

- Serial code
- Parallelization
- Why not parallelize the inner loop? to save overhead of repeated thread forks-joins
- Why must **j** be private? To avoid race condition among the threads

Nested parallelism

When a thread in a parallel region encounters another parallel construct, it may create a new team of threads and become the master of the new team.

Parallel tasks

Common mistakes

Race condition Deadlock

Not all computations are simple

Not all computations are simple loops where the data can be evenly divided among threads without any dependencies between threads

An example is finding the location and value of the largest element in an array

Not all computations are simple, competing threads

All threads are potentially accessing and changing the same values, **maxloc** and **maxval**.

1. OpenMP provides several ways to coordinate access to shared values

1. Only one thread at a time can execute the following statement (not block). We can use the critical option

1. Only one thread at a time can execute the following block

Atomic may be faster than critical but depends on hardware

How to find the max value using OpenMP

Write down the simplest algorithm and look carefully for race conditions. How would you handle them? The first step would be to parallelize as

Then deal with the race conditions

Write down the simplest algorithm and look carefully for race conditions. How would you handle them? The first step would be to parallelize as

Exercise: write a code which implements this and give an estimate on performance. Perform several runs, with a serial code only with and without vectorization and compare the serial code with the one that uses OpenMP. Run on different archictectures if you can.

What can slow down OpenMP performance?

Give it a thought!

What can slow down OpenMP performance?

Performance poor because we insisted on keeping track of the maxval and location during the execution of the loop.

- We do not care about the value during the execution of the loop, just the value at the end.

This is a common source of performance issues, namely the description of the method used to compute a value imposes additional, unnecessary requirements or properties

Idea: Have each thread find the maxloc in its own data, then combine and use temporary arrays indexed by thread number to hold the values found by each thread

Find the max location for each thread

Combine the values from each thread

Note that we let the master process perform the last operation.

Matrix-matrix multiplication

This code computes the norm of a vector using OpenMp

```
//  OpenMP program to compute vector norm by adding two other vectors
#include <cstdlib>
#include <iostream>
#include <cmath>
```

```
#include <iomanip>
#include  <omp.h>
# include <ctime>

using namespace std; // note use of namespace
int main (int argc, char* argv[])
{
  // read in dimension of vector
  int n = atoi(argv[1]);
  double *a, *b, *c;
  int i;
  int thread_num;
  double wtime, Norm2, s, angle;
  cout << "  Perform addition of two vectors and compute the norm-2." << endl;
  omp_set_num_threads(4);
  thread_num = omp_get_max_threads ();
  cout << "  The number of processors available = " << omp_get_num_procs () << endl ;
  cout << "  The number of threads available    = " << thread_num <<  endl;
  cout << "  The matrix order n                 = " << n << endl;

  s = 1.0/sqrt( (double) n);
  wtime = omp_get_wtime ( );
  // Allocate space for the vectors to be used
  a = new double [n]; b = new double [n]; c = new double [n];
  // Define parallel region
# pragma omp parallel for default(shared) private (angle, i) reduction(+:Norm2)
  // Set up values for vectors  a and b
  for (i = 0; i < n; i++){
      angle = 2.0*M_PI*i/ (( double ) n);
      a[i] = s*(sin(angle) + cos(angle));
      b[i] =  s*sin(2.0*angle);
      c[i] = 0.0;
  }
  // Then perform the vector addition
  for (i = 0; i < n; i++){
     c[i] += a[i]+b[i];
  }
  // Compute now the norm-2
  Norm2 = 0.0;
  for (i = 0; i < n; i++){
     Norm2  += c[i]*c[i];
  }
// end parallel region
  wtime = omp_get_wtime ( ) - wtime;
  cout << setiosflags(ios::showpoint | ios::uppercase);
  cout << setprecision(10) << setw(20) << "Time used  for norm-2 computation=" << wtime  << endl;
  cout << " Norm-2  = " << Norm2 << endl;
  // Free up space
  delete[] a;
  delete[] b;
  delete[] c;
  return 0;
}
```

Matrix-matrix multiplication

This the matrix-matrix multiplication code with plain c++ memory allocation using OpenMP

```
//  Matrix-matrix multiplication and Frobenius norm of a matrix with OpenMP
#include <cstdlib>
#include <iostream>
```

```
#include <cmath>
#include <iomanip>
#include  <omp.h>
# include <ctime>

using namespace std; // note use of namespace
int main (int argc, char* argv[])
{
  // read in dimension of square matrix
  int n = atoi(argv[1]);
  double **A, **B, **C;
  int i, j, k;
  int thread_num;
  double wtime, Fsum, s, angle;
  cout << "  Compute matrix product C = A * B and Frobenius norm." << endl;
  omp_set_num_threads(4);
  thread_num = omp_get_max_threads ();
  cout << "  The number of processors available = " << omp_get_num_procs () << endl ;
  cout << "  The number of threads available    = " << thread_num <<  endl;
  cout << "  The matrix order n                 = " << n << endl;

  s = 1.0/sqrt( (double) n);
  wtime = omp_get_wtime ( );
  // Allocate space for the two matrices
  A = new double*[n]; B = new double*[n]; C = new double*[n];
  for (i = 0; i < n; i++){
    A[i] = new double[n];
    B[i] = new double[n];
    C[i] = new double[n];
  }
  // Define parallel region
# pragma omp parallel for default(shared) private (angle, i, j, k) reduction(+:Fsum)
  // Set up values for matrix A and B and zero matrix C
  for (i = 0; i < n; i++){
    for (j = 0; j < n; j++) {
      angle = 2.0*M_PI*i*j/ (( double ) n);
      A[i][j] = s * ( sin ( angle ) + cos ( angle ) );
      B[j][i] =  A[i][j];
    }
  }
  // Then perform the matrix-matrix multiplication
  for (i = 0; i < n; i++){
    for (j = 0; j < n; j++) {
       C[i][j] =  0.0;
       for (k = 0; k < n; k++) {
            C[i][j] += A[i][k]*B[k][j];
       }
    }
  }
  // Compute now the Frobenius norm
  Fsum = 0.0;
  for (i = 0; i < n; i++){
    for (j = 0; j < n; j++) {
      Fsum += C[i][j]*C[i][j];
    }
  }
  Fsum = sqrt(Fsum);
// end parallel region and letting only one thread perform I/O
  wtime = omp_get_wtime ( ) - wtime;
  cout << setiosflags(ios::showpoint | ios::uppercase);
  cout << setprecision(10) << setw(20) << "Time used  for matrix-matrix multiplication=" << wtime  <<
  cout << "  Frobenius norm  = " << Fsum << endl;
  // Free up space
  for (int i = 0; i < n; i++){
```

```
    delete[] A[i];
    delete[] B[i];
    delete[] C[i];
  }
  delete[] A;
  delete[] B;
  delete[] C;
  return 0;
}
```

What is Message Passing Interface (MPI)?

MPI is a library, not a language. It specifies the names, calling sequences and results of functions or subroutines to be called from C/C++ or Fortran programs, and the classes and methods that make up the MPI C++ library. The programs that users write in Fortran, C or C++ are compiled with ordinary compilers and linked with the MPI library.

MPI programs should be able to run on all possible machines and run all MPI implementetations without change.

An MPI computation is a collection of processes communicating with messages.

Going Parallel with MPI

Task parallelism: the work of a global problem can be divided into a number of independent tasks, which rarely need to synchronize. Monte Carlo simulations or numerical integration are examples of this.

MPI is a message-passing library where all the routines have corresponding C/C++-binding and Fortran-binding (routine names are in uppercase, but can also be in lower case)

```
MPI_COMMAND_NAME
```

MPI is a library

MPI is a library specification for the message passing interface, proposed as a standard.

- independent of hardware;
- not a language or compiler specification;
- not a specific implementation or product.

A message passing standard for portability and ease-of-use. Designed for high performance.

Insert communication and synchronization functions where necessary.

Bindings to MPI routines

MPI is a message-passing library where all the routines have corresponding C/C++-binding and Fortran-binding (routine names are in uppercase, but can also be in lower case)

```
MPI_COMMAND_NAME
```

The discussion in these slides focuses on the C++ binding.

Communicator

- A group of MPI processes with a name (context).
- Any process is identified by its rank. The rank is only meaningful within a particular communicator.
- By default the communicator contains all the MPI processes.
- Mechanism to identify subset of processes.
- Promotes modular design of parallel libraries.

Some of the most important MPI functions

- MPI_Init - initiate an MPI computation
- $MPI\_Finalize$ - terminate the MPI computation and clean up
- MPI_Comm_size - how many processes participate in a given MPI communicator?
- MPI_Comm_rank - which one am I? (A number between 0 and size-1.)
- MPI_Send - send a message to a particular process within an MPI communicator
- MPI_Recv - receive a message from a particular process within an MPI communicator
- MPI_reduce or $MPI\_Allreduce$, send and receive messages

The first MPI C/C++ program

Let every process write "Hello world" (oh not this program again!!) on the standard output.

The Fortran program

```
PROGRAM hello
INCLUDE "mpif.h"
INTEGER:: size, my_rank, ierr

CALL  MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, size, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, my_rank, ierr)
WRITE(*,*)"Hello world, I've rank ",my_rank," out of ",size
CALL MPI_FINALIZE(ierr)

END PROGRAM hello
```

Note 1

- The output to screen is not ordered since all processes are trying to write to screen simultaneously.
- It is the operating system which opts for an ordering.
- If we wish to have an organized output, starting from the first process, we may rewrite our program as in the next example.

Ordered output with MPIBarrier

Note 2

- Here we have used the $MPI\_Barrier$ function to ensure that that every process has completed its set of instructions in a particular order.
- A barrier is a special collective operation that does not allow the processes to continue until all processes in the communicator (here MPI_COMM_WORLD) have called $MPI\_Barrier$.
- The barriers make sure that all processes have reached the same point in the code. Many of the collective operations like $MPI\_ALLREDUCE$ to be discussed later, have the same property; that is, no process can exit the operation until all processes have started.

However, this is slightly more time-consuming since the processes synchronize between themselves as many times as there are processes. In the next Hello world example we use the send and receive functions in order to a have a synchronized action.

Ordered output

```
.....
int numprocs, my_rank, flag;
MPI_Status status;
MPI_Init (&nargs, &args);
MPI_Comm_size (MPI_COMM_WORLD, &numprocs);
MPI_Comm_rank (MPI_COMM_WORLD, &my_rank);
if (my_rank > 0)
MPI_Recv (&flag, 1, MPI_INT, my_rank-1, 100,
          MPI_COMM_WORLD, &status);
cout << "Hello world, I have  rank " << my_rank << " out of "
<< numprocs << endl;
if (my_rank < numprocs-1)
MPI_Send (&my_rank, 1, MPI_INT, my_rank+1,
         100, MPI_COMM_WORLD);
MPI_Finalize ();
```

Note 3

The basic sending of messages is given by the function MPI_SEND, which in C/C++ is defined as This single command allows the passing of any kind of variable, even a large array, to any group of tasks. The variable **buf** is the variable we wish to send while **count** is the number of variables we are passing. If we are passing only a single value, this should be 1.

If we transfer an array, it is the overall size of the array. For example, if we want to send a 10 by 10 array, count would be $10 \times 10 = 100$ since we are actually passing 100 values.

Note 4

Once you have sent a message, you must receive it on another task. The function MPI_RECV is similar to the send call.

The arguments that are different from those in MPI_SEND are **buf** which is the name of the variable where you will be storing the received data, **source** which replaces the destination in the send command. This is the return ID of the sender.

Finally, we have used MPI_Status_status, where one can check if the receive was completed.

The output of this code is the same as the previous example, but now process 0 sends a message to process 1, which forwards it further to process 2, and so forth.

Numerical integration in parallel

Integrating π.

- The code example computes π using the trapezoidal rules.
- The trapezoidal rule

$$I = \int_a^b f(x)dx \approx h\left(f(a)/2 + f(a+h) + f(a+2h) + \cdots + f(b-h) + f(b)/2\right).$$

Click on this link for the full program.

Dissection of trapezoidal rule with MPI_reduce

Dissection of trapezoidal rule

Integrating with MPI

How do I use MPI_reduce?

Here we have used

The two variables *senddata* and *resultdata* are obvious, besides the fact that one sends the address of the variable or the first element of an array. If they are arrays they need to have the same size. The variable *count* represents the total dimensionality, 1 in case of just one variable, while $MPI\_Datatype$ defines the type of variable which is sent and received.

The new feature is MPI_Op. It defines the type of operation we want to do.

More on MPI_Reduce

In our case, since we are summing the rectangle contributions from every process we define $MPI\_Op = MPI\_SUM$. If we have an array or matrix we can search for the largest og smallest element by sending either MPI_MAX or MPI_MIN. If we want the location as well (which array element) we simply transfer MPI_MAXLOC or MPI_MINOC. If we want the product we write MPI_PROD.

$MPI\_Allreduce$ is defined as

Dissection of trapezoidal rule

We use MPI_reduce to collect data from each process. Note also the use of the function MPI_Wtime.

Dissection of trapezoidal rule