Introduction to High-Performance Scientific Computing

Victor Eijkhout
with
Edmond Chow, Robert van de Geijn

1st edition, 2010

Preface

The field of high performance scientific computing lies at the crossroads of a number of disciplines and skill sets, and correspondingly, for someone to be successful at using high performance computing in science requires at least elementary knowledge of and skills in all these areas. Computations stem from an application context, so some acquaintance with physics and engineering sciences is desirable. Then, problems in these application areas are typically translated into linear algebraic, and sometimes combinatorial, problems, so a computational scientist needs knowledge of several aspects of numerical analysis, linear algebra, and discrete mathematics. An efficient implementation of the practical formulations of the application problems requires some understanding of computer architecture, both on the CPU level and on the level of parallel computing. Finally, in addition to mastering all these sciences, a computational sciences needs some specific skills of software management.

While good texts exist on applied physics, numerical linear algebra, computer architecture, parallel computing, performance optimization, no book brings together these strands in a unified manner. The need for a book such as the present was especially apparent at the Texas Advanced Computing Center: users of the facilities there often turn out to miss crucial parts of the background that would make them efficient computational scientists. This book, then, comprises those topics that seem indispensible for scientists engaging in large-scale computations.

The contents of this book are a combination of theoretical material and self-guided tutorials on various practical skills. The theory chapters have exercises that can be assigned in a classroom, however, their placement in the text is such that a reader not inclined to do exercises can simply take them as statement of fact.

The tutorials should be done while sitting at a computer. Given the practice of scientific computing, they have a clear Unix bias.

Note This book is unfinished and open for comments. What is missing or incomplete or unclear? Is material presented in the wrong sequence? Kindly mail me with any comments you may have.

Victor Eijkhout `eijkhout@tacc.utexas.edu`
Research Scientist
Texas Advanced Computing Center
The University of Texas at Austin

Acknowledgement Helpful discussions with Kazushige Goto and John McCalpin are gratefully acknowledged. Thanks to Dan Stanzione for his notes on cloud computing and Ernie Chan for his notes on scheduling of block algorithms. Thanks to Elie de Brauwer for many comments.

Introduction

Scientific computing is the cross-disciplinary field at the intersection of modeling scientific processes, and the use of computers to produce quantitative results from these models. As a definition, we may posit

The efficient computation of constructive methods in applied mathematics.

This clearly indicates the three branches of science that scientific computing touches on:

- Applied mathematics: the mathematical modeling of real-world phenomena. Such modeling often leads to implicit descriptions, for instance in the form of partial differential equations. In order to obtain actual tangible results we need a constructive approach.
- Numerical analysis provides algorithmic thinking about scientific models. It offers a constructive approach to solving the implicit models, with an analysis of cost and stability.
- Computing takes numerical algorithms and analyzes the efficacy of implementing them on actually existing, rather than hypothetical, computing engines.

One might say that 'computing' became a scientific field in its own right, when the mathematics of real-world phenomena was asked to be constructive, that is, to go from proving the existence of solutions to actually obtaining them. At this point, algorithms become an object of study themselves, rather than a mere tool.

The study of algorithms became important when computers were invented. Since mathematical operations now were endowed with a definable time cost, complexity of algoriths became a field of study; since computing was no longer performed in 'real' numbers but in representations in finite bitstrings, the accuracy of algorithms needed to be studied. (Some of these considerations predate the existence of computers, having been inspired by computing with mechanical calculators.)

A prime concern in scientific computing is efficiency. While to some scientists the abstract fact of the existence of a solution is enough, in computing we actually want that solution, and preferably yesterday. For this reason, we will be quite specific about the efficiency of both algorithms and hardware.

Contents

Chapter 1

Sequential Computer Architecture

In order to write efficient scientific codes, it is important to understand computer architecture. The difference in speed between two codes that compute the same result can range from a few percent to orders of magnitude, depending only on factors relating to how well the algorithms are coded for the processor architecture. Clearly, it is not enough to have an algorithm and 'put it on the computer': some knowledge of computer architecture is advisable, sometimes crucial.

Some problems can be solved on a single CPU, others need a parallel computer that comprises more than one processor. We will go into detail on parallel computers in the next chapter, but even for parallel processing, it is necessary to understand the invidual CPUs.

In this chapter, we will focus on what goes on inside a CPU and its memory system. We start with a brief general discussion of how instructions are handled, then we will look into the arithmetic processing in the processor core; last but not least, we will devote much attention to the movement of data between memory and the processor, and inside the processor. This latter point is, maybe unexpectedly, very important, since memory access is typically much slower than executing the processor's instructions, making it the determining factor in a program's performance; the days when 'flop[1] counting' was the key to predicting a code's performance are long gone. This discrepancy is in fact a growing trend, so the issue of dealing with memory traffic has been becoming more important over time, rather than going away.

This chapter will give you a basic understanding of the issues involved in CPU design, how it affects performance, and how you can code for optimal performance. For much more detail, see an online book about PC architecture [55], and the standard work about computer architecture, Hennesey and Patterson [49].

1.1 The Von Neumann architecture

While computers, and most relevantly for this chapter, their processors, can differ in any number of details, they also have many aspects in common. On a very high level of abstraction, many architectures can be described as

1. Floating Point Operation.

von Neumann architectures. This describes a design with an undivided memory that stores both program and data ('stored program'), and a processing unit that executes the instructions, operating on the data.

This setup distinguishes modern processors for the very earliest, and some special purpose contemporary, designs where the program was hard-wired. It also allows programs to modify themselves or generate other programs, since instructions and data are in the same storage. This allows us to have editors and compilers: the computer treats program code as data to operate on. In this book we will not explicitly discuss compilers, the programs that translate high level languages to machine instructions. However, on occasion we will discuss how a program at high level can be written to ensure efficiency at the low level.

In scientific computing, however, we typically do not pay much attention to program code, focusing almost exclusively on data and how it is moved about during program execution. For most practical purposes it is as if program and data are stored separately. The little that is essential about instruction handling can be described as follows.

The machine instructions that a processor executes, as opposed to the higher level languages users write in, typically specify the name of an operation, as well as of the locations of the operands and the result. These locations are not expressed as memory locations, but as *registers*: a small number of named memory locations that are part of the CPU[2]. As an example, here is a simple C routine

```
void store(double *a, double *b, double *c) {
 *c = *a + *b;
}
```

and its X86 assembler output, obtained by[3] `gcc -O2 -S -o - store.c`:

```
        .text
        .p2align 4,,15
.globl store
        .type   store, @function
store:
        movsd   (%rdi), %xmm0 # Load *a to %xmm0
        addsd   (%rsi), %xmm0 # Load *b and add to %xmm0
        movsd   %xmm0, (%rdx) # Store to *c
        ret
```

The instructions here are:

- A load from memory to register;
- Another load, combined with an addition;
- Writing back the result to memory.

Each instruction is processed as follows:

2. Direct-to-memory architectures are rare, though they have existed. The Cyber 205 supercomputer in the 1980s could have 3 data streams, two from memory to the processor, and one back from the processor to memory, going on at the same time. Such an architecture is only feasible if memory can keep up with the processor speed, which is no longer the case these days.
3. This is 64-bit output; add the option `-m64` on 32-bit systems.

- Instruction fetch: the next instruction according to the *program counter* is loaded into the processor. We will ignore the questions of how and from where this happens.
- Instruction decode: the processor inspects the instruction to determine the operation and the operands.
- Memory fetch: if necessary, data is brought from memory into a register.
- Execution: the operation is executed, reading data from registers and writing it back to a register.
- Write-back: for store operations, the register contents is written back to memory.

Complicating this story, contemporary CPUs operate on several instructions simultaneously, which are said to be 'in flight', meaning that they are in various stages of completion. This is the basic idea of the *superscalar* CPU architecture, and is also referred to as *instruction-level parallelism*. Thus, while each instruction can take several clock cycles to complete, a processor can complete one instruction per cycle in favourable circumstances; in some cases more than one instruction can be finished per cycle.

The main statistic that is quoted about CPUs is their Gigahertz rating, implying that the speed of the processor is the main determining factor of a computer's performance. While speed obviously correlates with performance, the story is more complicated. Some algorithms are *cpu-bound*, and the speed of the processor is indeed the most important factor; other algorithms are *memory-bound*, and aspects such as bus speed and cache size, to be discussed later, become important.

In scientific computing, this second category is in fact quite prominent, so in this chapter we will devote plenty of attention to the process that moves data from memory to the processor, and we will devote relatively little attention to the actual processor.

1.1.1 Floating point units

Many modern processors are capable of doing multiple operations simultaneously, and this holds in particular for the arithmetic part. For instance, often there are separate addition and multiplication units; if the compiler can find addition and multiplication operations that are independent, it can schedule them so as to be executed simultaneously, thereby doubling the performance of the processor. In some cases, a processor will have multiple addition or multiplication units.

Another way to increase performance is to have a 'fused multiply-add' unit, which can execute the instruction $x \leftarrow ax + b$ in the same amount of time as a separate addition or multiplication. Together with pipelining (see below), this means that a processor has an asymptotic speed of several floating point operations per clock cycle.

1.1.1.1 Pipelining

The floating point add and multiply units of a processor are pipelined, which has the effect that a stream of independent operations can be performed at an asymptotic speed of one result per clock cycle.

The idea behind a pipeline is as follows. Assume that an operation consists of multiple simpler operations, and that for each suboperation there is separate hardware in the processor. For instance, an multiply instruction can have the following components:

- Decoding the instruction, including finding the locations of the operands.

Processor	floating point units	max operations per cycle
Pentium4, Opteron	2 add or 2 mul	2
Woodcrest, Barcelona	2 add + 2 mul	4
IBM POWER4, POWER5, POWER6	2 FMA	4
IBM BG/L, BG/P	1 SIMD FMA	4
SPARC IV	1 add + 1 mul	2
Itanium2	2 FMA	4

Table 1.1: Floating point capabilities of several current processor architectures

- Copying the operands into registers ('data fetch').
- Aligning the exponents; the multiplication $.3 \times 10^{-1} \times .2 \times 10^2$ becomes $.0003 \times 10^2 \times .2 \times 10^2$.
- Executing the multiplication, in this case giving $.00006 \times 10^2$.
- Normalizing the result, in this example to $.6 \times 10^0$.
- Storing the result.

These parts are often called the 'stages' or 'segments' of the pipeline.

If every component is designed to finish in 1 clock cycle, the whole instruction takes 6 cycles. However, if each has its own hardware, we can execute two multiplications in less than 12 cycles:

- Execute the decode stage for the first operation;
- Do the data fetch for the first operation, and at the same time the decode for the second.
- Execute the third stage for the first operation and the second stage of the second operation simultaneously.
- Et cetera.

You see that the first operation still takes 6 clock cycles, but the second one is finished a mere 1 cycle later. This idea can be extended to more than two operations: the first operation still takes the same amount of time as before, but after that one more result will be produced each cycle. Formally, executing n operations on a s-segment pipeline takes $s + n - 1$ cycles, as opposed to ns in the classical case.

Exercise 1.1. Let us compare the speed of a classical floating point unit, and a pipelined one. If the pipeline has s stages, what is the asymptotic speedup? That is, with $T_0(n)$ the time for n operations on a classical CPU, and $T_s(n)$ the time for n operations on an s-segment pipeline, what is $\lim_{n\to\infty}(T_0(n)/T_s(n))$?
Next you can wonder how long it takes to get close to the asymptotic behaviour. Define $S_s(n)$ as the speedup achieved on n operations. The quantity $n_{1/2}$ is defined as the value of n such that $S_s(n)$ is half the asymptotic speedup. Give an expression for $n_{1/2}$.

Since a vector processor works on a number of instructions simultaneously, these instructions have to be independent. The operation $\forall_i\colon a_i \leftarrow b_i + c_i$ has independent additions; the operation $\forall_i\colon a_{i+1} \leftarrow a_i b_i + c_i$ feeds the result of one iteration (a_i) to the input of the next ($a_{i+1} = \ldots$), so the operations are not independent.

A pipelined processor can speed up operations by a factor of $4, 5, 6$ with respect to earlier CPUs. Such numbers were typical in the 1980s when the first successful vector computers came on the market. These days, CPUs can have 20-stage pipelines. Does that mean they are incredibly fast? This question is a bit complicated. Chip designers

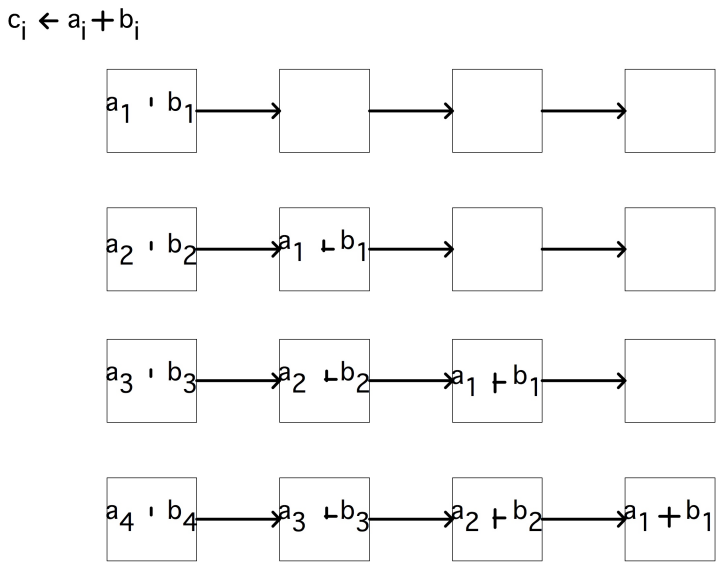


Figure 1.1: Schematic depiction of a pipelined operation

continue to increase the clock rate, and the pipeline segments can no longer finish their work in one cycle, so they are further spit up. Sometimes there are even segments in which nothing happens: that time is needed to make sure data can travel to a different part of the chip in time.

The amount of improvement you can get from a pipelined CPU is limited, so in a quest for ever higher performance several variations on the pipeline design have been tried. For instance, the Cyber 205 had separate addition and multiplication pipelines, and it was possible to feed one pipe into the next without data going back to memory first. Operations like $\forall_i\colon a_i \leftarrow b_i + c \cdot d_i$ were called 'linked triads' (because of the number of paths to memory, one input operand had to be scalar).

Exercise 1.2. Analyse the speedup and $n_{1/2}$ of linked triads.

Another way to increase performance is to have multiple identical pipes. This design was perfected by the NEC SX series. With, for instance, 4 pipes, the operation $\forall_i\colon a_i \leftarrow b_i + c_i$ would be split module 4, so that the first pipe operated on indices $i = 4 \cdot j$, the second on $i = 4 \cdot j + 1$, et cetera.

Exercise 1.3. Analyze the speedup and $n_{1/2}$ of a processor with multiple pipelines that operate in parallel. That is, suppose that there are p independent pipelines, executing the same instruction, that can each handle a stream of operands.

(The reason we are mentioning some fairly old computers here is that true pipeline supercomputers hardly exist anymore. In the US, the Cray X1 was the last of that line, and in Japan only NEC still makes them. However, the functional units of a CPU these days are pipelined, so the notion is still important.)

Exercise 1.4. The operation

```
for (i) {
  x[i+1] = a[i]*x[i] + b[i];
}
```

can not be handled by a pipeline or SIMD processor because there is a *dependency* between input of one iteration of the operation and the output of the previous. However, you can transform the loop into one that is mathematically equivalent, and potentially more efficient to compute. Derive an expression that computes `x[i+2]` from `x[i]` without involving `x[i+1]`. This is known as *recursive doubling*. Assume you have plenty of temporary storage. You can now perform the calculation by

- Doing some preliminary calculations;
- computing `x[i],x[i+2],x[i+4],...`, and from these,
- compute the missing terms `x[i+1],x[i+3],...`.

Analyze the efficiency of this scheme by giving formulas for $T_0(n)$ and $T_s(n)$. Can you think of an argument why the preliminary calculations may be of lesser importance in some circumstances?

1.1.1.2 Peak performance

For marketing purposes, it may be desirable to define a 'top speed' for a CPU. Since a pipelined floating point unit can yield one result per cycle asymptotically, you would calculate the theoretical *peak performance* as the product of the clock speed (in ticks per second), number of floating point units, and the number of cores (see section 1.3). This top speed is unobtainable in practice, and very few codes come even close to it; see section 2.9. Later in this chapter you will learn the reasons that it is so hard to get this perfect performance.

1.1.1.3 Pipelining beyond arithmetic: instruction-level parallelism

In fact, nowadays, the whole CPU is pipelined. Not only floating point operations, but any sort of instruction will be put in the *instruction pipeline* as soon as possible. Note that this pipeline is no longer limited to identical instructions: the notion of pipeline is now generalized to any stream of partially executed instructions that are simultaneously "in flight".

This concept is also known as *instruction-level parallelism*, and it is facilitated by various mechanisms:

- multiple-issue: instructions that are independent can be started at the same time;
- pipelining: already mentioned, arithmetic units can deal with multiple operations in various stages of completion;
- branch prediction and speculative execution: a compiler can 'guess' whether a conditional instruction will evaluate to true, and execute those instructions accordingly;
- out-of-order execution: instructions can be rearranged if they are not dependent on each other, and if the resulting execution will be more efficient;
- prefetching: data can be speculatively requested before any instruction needing it is actually encountered (this is discussed further in section 1.2.5).

As clock frequency has gone up, the processor pipeline has grown in length to make the segments executable in less time. You have already seen that longer pipelines have a larger $n_{1/2}$, so more independent instructions are needed to make the pipeline run at full efficiency. As the limits to instruction-level parallelism are reached, making pipelines longer (sometimes called 'deeper') no longer pays off. This is generally seen as the reason that chip designers have moved to *multi-core* architectures as a way of more efficiently using the transistors on a chip; section 1.3.

There is a second problem with these longer pipelines: if the code comes to a branch point (a conditional or the test in a loop), it is not clear what the next instruction to execute is. At that point the pipeline can stall. CPUs have taken to *speculative execution'* for instance, by always assuming that the test will turn out true. If the code then takes the other branch (this is called a *branch misprediction*), the pipeline has to be cleared and restarted. The resulting delay in the execution stream is called the *branch penalty*.

1.1.1.4 8-bit, 16-bit, 32-bit, 64-bit

Processors are often characterized in terms of how big a chunk of data they can process as a unit. This can relate to

- The width of the path between processor and memory: can a 64-bit floating point number be loaded in one cycle, or does it arrive in pieces at the processor.
- The way memory is addressed: if addresses are limited to 16 bits, only 64,000 bytes can be identified. Early PCs had a complicated scheme with segments to get around this limitation: an address was specified with a segment number and an offset inside the segment.
- The number of bits in a register, in particular the size of the integer registers which manipulate data address; see the previous point. (Floating point register are often larger, for instance 80 bits in the x86 architecture.) This also corresponds to the size of a chunk of data that a processor can operate on simultaneously.
- The size of a floating point number. If the arithmetic unit of a CPU is designed to multiply 8-byte numbers efficiently ('double precision'; see section 3.2) then numbers half that size ('single precision') can sometimes be processed at higher efficiency, and for larger numbers ('quadruple precision') some complicated scheme is needed. For instance, a quad precision number could be emulated by two double precision numbers with a fixed difference between the exponents.

These measurements are not necessarily identical. For instance, the original Pentium processor had 64-bit data busses, but a 32-bit processor. On the other hand, the Motorola 68000 processor (of the original Apple Macintosh) had a 32-bit CPU, but 16-bit data busses.

The first Intel microprocessor, the 4004, was a 4-bit processor in the sense that it processed 4 bit chunks. These days, processors are 32-bit, and 64-bit is becoming more popular.

1.2 Memory Hierarchies

We will now refine the picture of the Von Neuman architecture, in which data is loaded immediately from memory to the processors, where it is operated on. This picture is unrealistic because of the so-called *memory wall*: the

memory is too slow to load data into the process at the rate the processor can absorb it. Specifically, a single load can take 1000 cycles, while a processor can perform several operations per cycle. (After this long wait for a load, the next load can come faster, but still too slow for the processor. This matter of wait time versus throughput will be addressed below in section 1.2.2.)

In reality, there will be various memory levels in between the floating point unit and the main memory: the registers and the caches. Each of these will be faster to a degree than main memory; unfortunately, the faster the memory on a certain level, the smaller it will be. This leads to interesting programming problems, which we will discuss in the rest of this chapter, and particularly section 1.5.

The use of registers is the first instance you will see of measures taken to counter the fact that loading data from memory is slow. Access to data in registers, which are built into the processor, is almost instantaneous, unlike main memory, where hundreds of clock cycles can pass between requesting a data item, and it being available to the processor.

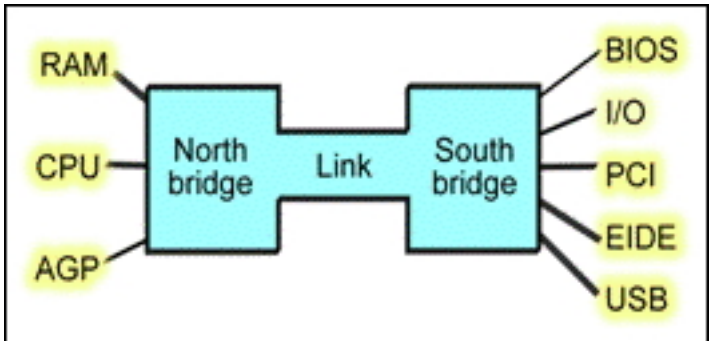
One advantage of having registers is that data can be kept in them during a computation, which obviates the need for repeated slow memory loads and stores. For example, in

```
s = 0;
for (i=0; i<n; i++)
  s += a[i]*b[i];
```

the variable `s` only needs to be stored to main memory at the end of the loop; all intermediate values are almost immediately overwritten. Because of this, a compiler will keep `s` in a register, eliminating the delay that would result from continuously storing and loading the variable. Such questions of *data reuse* will be discussed in more detail below; we will first discuss the components of the memory hierarchy.

1.2.1 Busses

The wires that move data around in a computer, from memory to cpu or to a disc controller or screen, are called *busses*. The most important one for us is the *Front-Side Bus (FSB)* which connects the processor to memory. In one popular architecture, this is called the 'north bridge', as opposed to the 'south bridge' which connects to external devices, with the exception of the graphics controller.



The bus is typically much slower than the processor, operating between 500MHz and 1GHz, which is the main reason that caches are needed. Besides the frequency, the bandwidth of a bus is also determined by the number of bits that can be moved per clock cycle. This is typically 64 or 128 in current architectures.

1.2.2 Latency and Bandwidth

Above, we mentioned in very general terms that accessing data in registers is almost instantaneous, whereas loading data from memory into the registers, a necessary step before any operation, incurs a substantial delay. We will now make this story slightly more precise.

There are two important concepts to describe the movement of data: *latency* and *bandwidth*. The assumption here is that requesting an item of data incurs an initial delay; if this item was the first in a stream of data, usually a consecutive range of memory addresses, the remainder of the stream will arrive with no further delay at a regular amount per time period.

Latency is the delay between the processor issuing a request for a memory item, and the item actually arriving. We can distinguish between various latencies, such as the transfer from memory to cache, cache to register, or summarize them all into the latency between memory and processor. Latency is measured in (nano) seconds, or clock periods.

If a processor executes instructions in the order they are found in the assembly code, then execution will often *stall* while data is being fetched from memory; this is also called *memory stall*. For this reason, a low latency is very important. In practice, many processors have 'out-of-order execution' of instructions, allowing them to perform other operations while waiting for the requested data. Programmers can take this into account, and code in a way that achieves *latency hiding*. Graphics Processing Units (GPUs) can switch very quickly between threads in order to achieve latency hiding.

Bandwidth is the rate at which data arrives at its destination, after the initial latency is overcome. Bandwidth is measured in bytes (kilobyes, megabytes, gigabyes) per second or per clock cycle. The bandwidth between two memory levels is usually the product of the cycle speed of the channel (the *bus speed*) and width of the bus: the number of bits that can be sent simultaneously in every cycle of the bus clock.

The concepts of latency and bandwidth are often combined in a formula for the time that a message takes from start to finish:

$$T(n) = \alpha + \beta n$$

where α is the latency and β is the inverse of the bandwidth: the time per byte.

Typically, the further away from the processor one gets, the longer the latency is, and the lower the bandwidth. These two factors make it important to program in such a way that, if at all possible, the processor uses data from cache or register, rather than from main memory. To illustrate that this is a serious matter, consider a vector addition

```
for (i)
  a[i] = b[i]+c[i]
```

Each iteration performs one floating point operation, which modern CPUs can do in one clock cycle by using pipelines. However, each iteration needs two numbers loaded and one written, for a total of 24 bytes[4] of memory

4. Actually, `a[i]` is loaded before it can be written, so there are 4 memory access, with a total of 32 bytes, per iteration.

traffic. Typical memory bandwidth figures (see for instance figure 1.2) are nowhere near 24 (or 32) bytes per cycle. This means that, without caches, algorithm performance can be bounded by memory performance.

The concepts of latency and bandwidth will also appear in parallel computers, when we talk about sending data from one processor to the next.

1.2.3 Registers

The registers are the memory level that is closest to the processor: any operation acts on data in register and leaves its result in register. Programs written in assembly language explicitly use these registers:

```
addl %eax, %edx
```

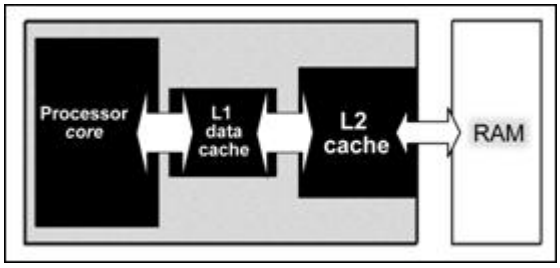
which is an instruction to add the content of one register to another. As you see in this sample instruction, registers are not numbered in memory, but have distinct names that are referred to in the assembly instruction.

Registers have a high bandwidth and low latency, because they are part of the processor. That also makes them a very scarce resource.

1.2.4 Caches

In between the registers, which contain the data that the processor operates on, and the main memory where lots of data can reside for a long time, are various levels of *cache* memory, that have lower latency and higher bandwidth than main memory. Data from memory travels the cache hierarchy to wind up in registers. The advantage to having cache memory is that if a data item is reused shortly after it was first needed, it will still be in cache, and therefore it can be accessed much faster than if it would have to be brought in from memory.

The caches are called 'level 1' and 'level 2' (or, for short, L1 and L2) cache; some processors can have an L3 cache. The L1 and L2 caches are part of the *die*, the processor chip, although for the L2 cache that is a recent development; the L3 cache is off-chip. The L1 cache is small, typically around 16Kbyte. Level 2 (and, when present, level 3) cache is more plentiful, up to several megabytes, but it is also slower. Unlike main memory, which is expandable, caches are fixed in size. If a version of a processor chip exists with a larger cache, it is usually considerably more expensive.



Data that is needed in some operation gets copied into the various caches on its way to the processor. If, some instructions later, a data item is needed again, it is first searched for in the L1 cache; if it is not found there, it is searched for in the L2 cache; if it is not found there, it is loaded from main memory. Finding data in cache is called a *cache hit*, and not finding it a *cache miss*.

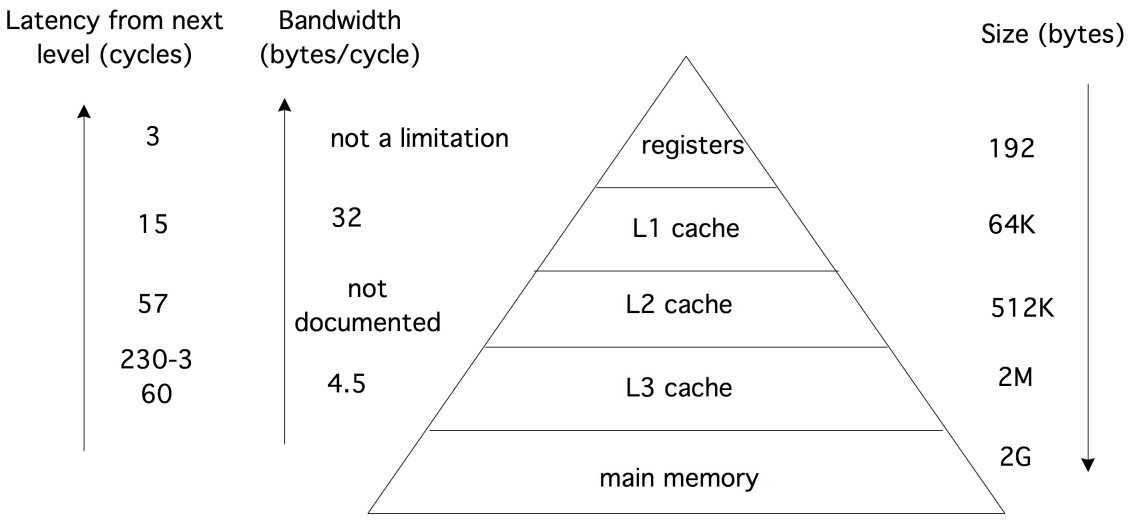


Figure 1.2: Memory hierarchy of an AMD Xeon, characterized by speed and size.

Figure 1.2 illustrates the basic facts of caches, in this case for the AMD Opteron chip: the closer caches are to the floating point units, the faster, but also the smaller they are. Some points about this figure.

- Loading data from registers is so fast that it does not constitute a limitation on algorithm execution speed. On the other hand, there are few registers. The Opteron[5] has 16 general purpose registers, 8 media and floating point registers, and 16 SIMD registers.
- The L1 cache is small, but sustains a bandwidth of 32 bytes, that is 4 double precision number, per cycle. This is enough to load two operands each for two operations, but note that the Opteron can actually perform 4 operations per cycle. Thus, to achieve peak speed, certain operands need to stay in register. The latency from L1 cache is around 3 cycles.
- The bandwidth from L2 and L3 cache is not documented and hard to measure due to cache policies (see below). Latencies are around 15 cycles for L2 and 50 for L3.
- Main memory access has a latency of more than 100 cycles, and a bandwidth of 4.5 bytes per cycle, which is about $1/7$th of the L1 bandwidth. However, this bandwidth is shared by the 4 cores of the opteron chip, so effectively the bandwidth is a quarter of this number. In a machine like Ranger, which has 4 chips per node, some bandwidth is spent on maintaining *cache coherence* (see section 1.3) reducing the bandwidth for each chip again by half.

On level 1, there are separate caches for instructions and data; the L2 and L3 cache contain both data and instructions.

You see that the larger caches are increasingly unable to supply data to the processors fast enough. For this reason it is necessary to code in such a way that data is kept as much as possible in the highest cache level possible. We will discuss this issue in detail in the rest of this chapter.

5. Specifically the server chip used in the Ranger supercomputer; desktop versions may have different specifications.

Exercise 1.5. The L1 cache is smaller than the L2 cache, and if there is an L3, the L2 is smaller than the L3. Give a practical and a theoretical reason why this is so.

1.2.4.1 Reuse is the name of the game

The presence of one or more caches is not immediately a guarantee for high performance: this largely depends on the *memory access pattern* of the code, and how well this exploits the caches. The first time that an item is referenced, it is copied from memory into cache, and through to the processor registers. The latency and bandwidth for this are not mitigated in any way by the presence of a cache. When the same item is referenced a second time, it may be found in cache, at a considerably reduced cost in terms of latency and bandwidth: caches have shorter latency and higher bandwidth than main memory.

We conclude that, first, an algorithm has to have an opportunity for data reuse. If every data item is used only once (as in addition of two vectors), there can be no reuse, and the presence of caches is largely irrelevant. A code will only benefit from the increased bandwidth and reduced latency of a cache if items in cache are referenced more than once; see section 1.4.1 for a detailed discussion.. An example would be the matrix-vector multiplication $y = Ax$ where each element of x is used in n operations, where n is the matrix dimension.

Secondly, an algorithm may theoretically have an opportunity for reuse, but it needs to be coded in such a way that the reuse is actually exposed. We will address these points in section 1.4.2. This second point especially is not trivial.

Some problems are small enough that they fit completely in cache, at least in the L3 cache. This is something to watch out for when *benchmarking*, since it gives a too rosy picture of processor performance.

1.2.4.2 Replacement policies

Data in cache and registers is placed there by the system, outside of programmer control. Likewise, the system decides when to overwrite data in the cache or in registers if it is not referenced in a while, and as other data needs to be placed there. Below, we will go into detail on how caches do this, but as a general principle, a Least Recently Used (LRU) cache replacement policy is used: if a cache is full and new data needs to be placed into it, the data that was least recently used is *flushed*, meaning that it is overwritten with the new item, and therefore no longer accessible. LRU is by far the most common replacement policy; other possibilities are FIFO (first in first out) or random replacement.

Exercise 1.6. Sketch a simple scenario, and give some (pseudo) code, to argue that LRU is preferable over FIFO as a replacement strategy.

1.2.4.3 Cache lines

Data is not moved between memory and cache, or between caches, in single bytes. Instead, the smallest unit of data moved is called a *cache line*. A typical cache line is 64 or 128 bytes long, which in the context of scientific computing implies 8 or 16 double precision floating point numbers. The cache line size for data moved into L2 cache can be larger than for data moved into L1 cache.

It is important to acknowledge the existence of cache lines in coding, since any memory access costs the transfer of several words (see section 1.5.3 for some practical discussion). An efficient program then tries to use the other items on the cache line, since access to them is effectively free. This phenomenon is visible in code that accesses arrays by *stride*: elements are read or written at regular intervals.

Stride 1 corresponds to sequential access of an array:

```
for (i=0; i<N; i++)
  ... = ... x[i] ...
```

A larger stride

```
for (i=0; i<N; i+=stride)
  ... = ... x[i] ...
```

implies that in every cache line only certain elements are used; a stride of 4 with a cache line size of 4 double precision numbers implies that in every cache line only one number is used, and therefore that $3/4$ of the used bandwidth has been wasted.

Some applications naturally lead to strides greater than 1, for instance, accessing only the real parts of an array of complex numbers; section 3.4.4. Also, methods that use *recursive doubling* often have a code structure that exhibits non-unit strides

```
for (i=0; i<N/2; i++)
  x[i] = y[2*i];
```

In this discussion of cachelines, we have implicitly assumed the beginning of a cacheline is also the beginning of a word, be that an integer or a floating point number. This need not be true: an 8-byte floating point number can be placed straddling the boundary between two cachelines. You can image that this is not good for performance. To force allocated space to be *aligned* with a cacheline boundary, you can do the following:

```
double *a;
a = malloc( /* some number of bytes */ +8 );
if ( (int)a % 8 != 0 ) { /* it is not 8-byte aligned */
  a += 1; /* advance address by 8 bytes, then */
  /* either: */
  a = ( (a>>3) <<3 );
  /* or: */
  a = 8 * ( ( (int)a )/8 );
}
```

This code allocates a block of memory, and, if necessary, shfits it right to have a starting address that is a multiple of 8. This sort of alignment can sometimes be forced by compiler options.

1.2.4.4 Cache mapping

Caches get faster, but also smaller, the closer to the floating point units they get; even the largest cache is considerably smaller than the main memory size. We already noted that this has implications for the cache replacement

strategy. Another issue we need to address in this context is that of *cache mapping*, which is the question of 'if an item is placed in cache, where does it get placed'. This problem is generally addressed by mapping the (main memory) address of the item to an address in cache, leading to the question 'what if two items get mapped to the same address'.

1.2.4.5 Direct mapped caches

The simplest cache mapping strategy is *direct mapping*. Suppose that memory addresses are 32 bits long, so that they can address 4G bytes; suppose further that the cache has 8K words, that is, 64k bytes, needing 16 bits to address. Direct mapping then takes from each memory address the last ('least significant') 16 bits, and uses these as the address of the data item in cache.

Direct mapping is very efficient because of its address calculations can be performed very quickly, leading to low latency, but it has a problem in practical applications. If two items are addressed that are separated by 8K words, they will be mapped to the same cache location, which will make certain calculations inefficient. Example:

```
double A[3][8192];
for (i=0; i<512; i++)
  a[2][i] = ( a[0][i]+a[1][i] )/2.;
```

or in Fortran:

```
real*8 A(8192,3);
do i=1,512
  a(i,3) = ( a(i,1)+a(i,2) )/2
end do
```

Here, the locations of `a[0][i]`, `a[1][i]`, and `a[2][i]` (or `a(i,1)`, `a(i,2)`, `a(i,3)`) are 8K from each other for every `i`, so the last 16 bits of their addresses will be the same, and hence they will be mapped to the same location in cache. The execution of the loop will now go as follows:

- The data at `a[0][0]` is brought into cache and register. This engenders a certain amount of latency. Together with this element, a whole cache line is transferred.
- The data at `a[1][0]` is brought into cache (and register, as we will not remark anymore from now on), together with its whole cache line, at cost of some latency. Since this cache line is mapped to the same location as the first, the first cache line is overwritten.
- In order to write the output, the cache line containing `a[2][0]` is brought into memory. This is again mapped to the same location, causing flushing of the cache line just loaded for `a[1][0]`.
- In the next iteration, `a[0][1]` is needed, which is on the same cache line as `a[0][0]`. However, this cache line has been flushed, so it needs to be brought in anew from main memory or a deeper cache level. In doing so, it overwrites the cache line that holds `a[1][0]`.
- A similar story hold for `a[1][1]`: it is on the cache line of a[1][0], which unfortunately has been overwritten in the previous step.

If a cache line holds four words, we see that each four iterations of the loop involve eight transfers of elements of `a`, where two would have sufficed, if it were not for the cache conflicts.

Exercise 1.7. In the example of direct mapped caches, mapping from memory to cache was done by using the final 16 bits of a 32 bit memory address as cache address. Show that the problems in this example go away if the mapping is done by using the first ('most significant') 16 bits as the cache address. Why is this not a good solution in general?

1.2.4.6 Associative caches

The problem of cache conflicts, outlined in the previous section, would be solved if any data item could go to any cache location. In that case there would be no conflicts, other than the cache filling up, in which case a cache replacement policy (section 1.2.4.2) would flush data to make room for the incoming item. Such a cache is called *fully associative*, and while it seems optimal, it is also very costly to build, and much slower in use than a direct mapped cache.

For this reason, the most common solution is to have a k-way *associative cache*, where k is at least two. In this case, a data item can go to any of k cache locations. Code would have to have a $k+1$-way conflict before data would be flushed prematurely as in the above example. In that example, a value of $k = 2$ would suffice, but in practice higher values are often encountered.

For instance, the Intel Woodcrest processor has

- an L1 cache of 32K bytes, that is 8-way set associative with a 64 byte cache line size;
- an L2 cache of 4M bytes, that is 8-way set associative with a 64 byte cache line size.

On the other hand, the AMD Barcelona chip has 2-way associativity for the L1 cache, and 8-way for the L2. A higher associativity ('way-ness') is obviously desirable, but makes a processor slower, since determining whether an address is already in cache becomes more complicated. For this reason, the associativity of the L1 cache, where speed is of the greatest importance, is typically lower than of the L2.

Exercise 1.8. Write a small cache simulator in your favourite language. Assume a k-way associative cache of 32 entries and an architecture with 16 bit addresses. Run the following experiment for $k = 1, 2, 4$:

1. Let k be the associativity of the simulated cache.
2. Write the translation from 16 bit address to $32/2^k$ bit cache address.
3. Generate 32 random machine addresses, and simulate storing them in cache.

Since the cache has 32 entries, optimally the 32 addresses can all be stored in cache. The chance of this actually happening is small, and often the data of one address will be removed from the cache when it conflicts with another address. Record how many addresses, out of 32, are actually stored in the cache. Do step 3 100 times, and plot the results; give median and average value, and the standard deviation. Observe that increasing the associativity improves the number of addresses stored. What is the limit behaviour? (For bonus points, do a formal statistical analysis.)

1.2.5 Prefetch streams

In the traditional von Neumann model (section 1.1), each instruction contains the location of its operands, so a CPU implementing this model would make a separate request for each new operand. In practice, often subsequent data items are adjacent or regularly spaced in memory. The memory system can try to detect such data patterns by looking at cache miss points, and request a *prefetch data stream*.

In its simplest form, the CPU will detect that consecutive loads come from two consecutive cache lines, and automatically issue a request for the next following cache line. This process can be repeated or extended if the code makes an actual request for that third cache line. Since these cache lines are now brought from memory well before they are needed, prefetch has the possibility of eliminating the latency for all but the first couple of data items.

The concept of *cache miss* now needs to be revisited a little. From a performance point of view we are only interested in *stall*s on cache misses, that is, the case where the computation has to wait for the data to be brought in. Data that is not in cache, but can be brought in while other instructions are still being processed, is not a problem. If an 'L1 miss' is understood to be only a 'stall on miss', then the term 'L1 cache refill' is used to describe all cacheline loads, whether the processor is stalling on them or not.

Since prefetch is controlled by the hardware, it is also described as *hardware prefetch*. Prefetch streams can sometimes be controlled from software, though often it takes assembly code to do so.

1.2.6 Memory banks

Above, we discussed issues relating to bandwdith. You saw that memory, and to a lesser extent caches, have a bandwidth that is less than what a processor can maximally absorb. The situation is actually even worse than the above discussion made it seem. For this reason, memory is often divided into *memory banks* that are interleaved: with four memory banks, words $0, 4, 8, \ldots$ are in bank 0, words $1, 5, 9, \ldots$ are in bank 1, et cetera.

Suppose we now access memory sequentially, then such 4-way interleaved memory can sustain four times the bandwidth of a single memory bank. Unfortunately, accessing by stride 2 will halve the bandwidth, and larger strides are even worse. In practice the number of memory banks will be higher, so that strided memory access with small strides will still have the full advertised bandwidth.

This concept of banks can also apply to caches. For instance, the cache lines in the L1 cache of the AMD Barcelona chip are 16 words long, divided into two interleaved banks of 8 words. This means that sequential access to the elements of a cache line is efficient, but strided access suffers from a deteriorated performance.

1.2.7 TLB and virtual memory

All of a program's data may not be in memory simultaneously. This can happen for a number of reasons:

- The computer serves multiple users, so the memory is not dedicated to any one user;
- The computer is running multiple programs, which together need more than the physically available memory;
- One single program can use more data than the available memory.

For this reason, computers use *Virtual memory*: if more memory is needed than is available, certain blocks of memory are written to disc. In effect, the disc acts as an extension of the real memory. This means that a block of data can be anywhere in memory, and in fact, if it is *swapped* in and out, it can be in different locations at different times. Swapping does not act on individual memory locations, but rather on *memory pages*: contiguous blocks of memory, from a few kilobytes to megabytes in size. (In an earlier generation of operating systems, moving memory to disc was a programmer's responsibility. Pages that would replace each other were called *overlays*.)

For this reason, we need a translation mechanism from the memory addresses that the program uses to the actual addresses in memory, and this translation has to be dynamic. A program has a 'logical data space' (typically starting from address zero) of the addresses used in the compiled code, and this needs to be translated during program execution to actual memory addresses. For this reason, there is a *page table* that specifies which memory pages contain which logical pages.

However, address translation by lookup in this table is slow, so CPUs have a *Translation Look-aside Buffer (TLB)*. The TLB is a cache of frequently used Page Table Entries: it provides fast address translation for a number of pages. If a program needs a memory location, the TLB is consulted to see whether this location is in fact on a page that is remembered in the TLB. If this is the case, the logical address is translated to a physical one; this is a very fast process. The case where the page is not remembered in the TLB is called a *TLB miss*, and the page lookup table is then consulted, if necessary bringing the needed page into memory. The TLB is (sometimes fully) associative (section 1.2.4.6), using an LRU policy (section 1.2.4.2).

A typical TLB has between 64 and 512 entries. If a program accesses data sequentially, it will typically alternate between just a few pages, and there will be no TLB misses. On the other hand, a program that access many random memory locations can experience a slowdown because of such misses.

Section 1.5.4 and appendix D.5 discuss some simple code illustrating the behaviour of the TLB.

[There are some complications to this story. For instance, there is usually more than one TLB. The first one is associated with the L2 cache, the second one with the L1. In the AMD Opteron, the L1 TLB has 48 entries, and is is fully (48-way) associative, while the L2 TLB has 512 entries, but is only 4-way associative. This means that there can actually be TLB conflicts. In the discussion above, we have only talked about the L2 TLB. The reason that this can be associated with the L2 cache, rather than with main memory, is that the translation from memory to L2 cache is deterministic.]

1.3 Multi-core chips

In recent years, the limits of performance have been reached for the traditional processor chip design.

- Clock frequency can not increased further, since it increases energy consumption, heating the chips too much. Figure 1.3 gives a dramatic illustration of the heat that a chip would give off, if single-processor trends had continued. The reason for this is that the power dissipation of a chip is propertional to the voltage squared times the frequency. Since voltage and frequency are proportional, that makes power proportional to the third power of the frequency.

- It is not possible to extract more instruction-level parallelism from codes, either because of compiler limitations, because of the limited amount of intrinsically available parallelism, or because branch prediction makes it impossible (see section 1.1.1.3).

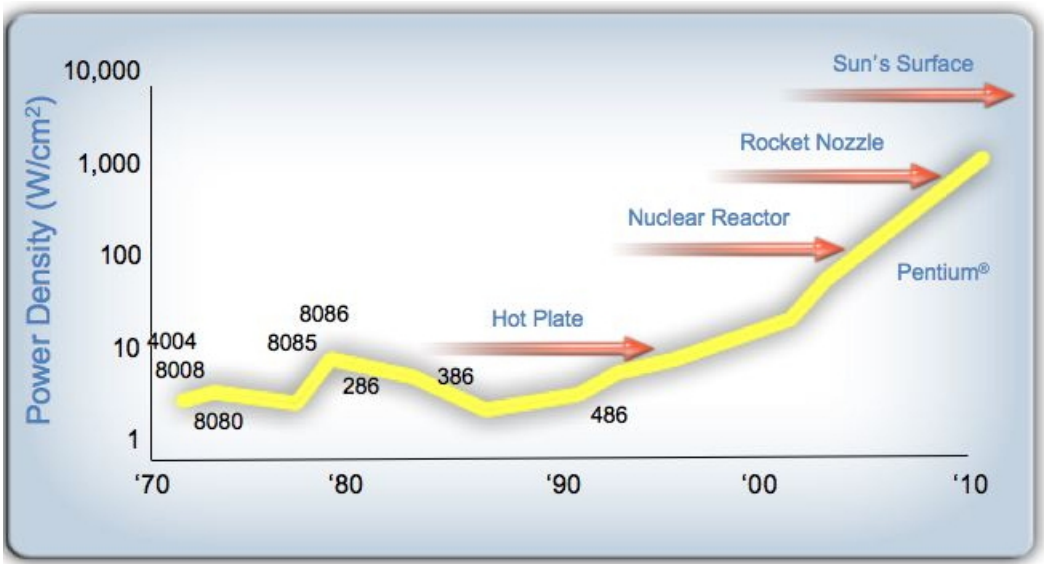


Figure 1.3: Projected heat dissipation of a CPU if trends had continued – this graph courtesy Pat Helsinger

One of the ways of getting a higher utilization out of a single processor chip is then to move from a strategy of further sophistication of the single processor, to a division of the chip into multiple processing 'cores'[6]. The separate cores can work on unrelated tasks, or, by introducing what is in effect data parallelism (section 2.2.1), collaborate on a common task at a higher overall efficiency.

While first multi-core chips were simply two processors on the same die, later generations incorporated L2 caches that were shared between the two processor cores. This design makes it efficient for the cores to work jointly on the same problem. The cores would still have their own L1 cache, and these separate caches lead to a *cache coherence* problem; see section 1.3.1 below.

We note that the term 'processor' is now ambiguous: it can refer to either the chip, or the processor core on the chip. For this reason, we mostly talk about a *socket* for the whole chip and *core* for part containing one arithmetic and logic unit and having its own registers. Currently, CPUs with 4 or 6 cores are on the market and 8-core chips will be available shortly. The core count is likely to go up in the future: Intel has already shown an 80-core prototype that is developed into the 48 core 'Single-chip Cloud Computer', illustrated in fig 1.4. This chip has a structure with 24 dual-core 'tiles' that are connected through a 2D mesh network. Only certain tiles are connected to a memory controller, others can not reach memory other than through the on-chip network.

6. Another solution is Intel's *Hyperthreading*, which lets a processor mix the instructions of several instruction streams. The benefits of this are strongly dependent on the individual case. However, this same mechanism is exploited with great success in GPUs; see section **??**.

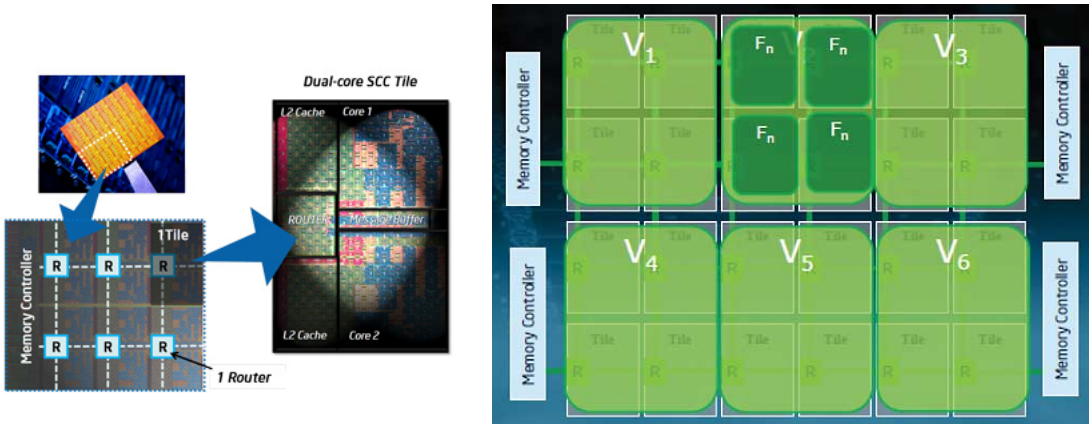


Figure 1.4: Structure of the Intel Single-chip Cloud Computer chip

With this mix of shared and private caches, the programming model for multi-core processors is becoming a hybrid between shared and distributed memory:

Core The cores have their own private L1 cache, which is a sort of distributed memory. The above mentioned Intel 80-core prototype has the cores communicating in a distributed memory fashion.

Socket On one socket, there is often a shared L2 cache, which is shared memory for the cores.

Node There can be multiple sockets on a single 'node' or motherboard, accessing the same shared memory.

Network Distributed memory programming (see the next chapter) is needed to let nodes communicate.

1.3.1 Cache coherence

With parallel processing, there is the potential for a conflict if more than one processor has a copy of the same data item. The problem of ensuring that all cached data are an accurate copy of main memory, is referred to as *cache coherence*: if one processor alters its copy, how is the other copy updated.

In distributed memory architectures, a dataset is usually partitioned disjointly over the processors, so conflicting copies of data can only arise with knowledge of the user, and it is up to the user to deal with the problem. The case of shared memory is more subtle: since processes access the same main memory, it would seem that conflicts are in fact impossible. However, processor typically have some private cache, which contains copies of data from memory, so conflicting copies can occur. This situation arises in particular in multi-core designs.

Suppose that two cores have a copy of the same data item in their (private) L1 cache, and one modifies its copy. Now the other has cached data that is no longer an accurate copy of its counterpart, so it needs to reload that item. This is known as *maintaining cache coherence*, and it is done on a very low level of the processor, with no programmer involvement needed. However, it will slow down the computation, and it wastes bandwidth to the core that could otherwise be used for loading or storing operands.

The state of a cache line with respect to a data item in main memory is usually described as one of the following:

Scratch: the cache line does not contain a copy of the item;
Valid: the cache line is a correct copy of data in main memory;
Reserved: the cache line is the *only* copy of that piece of data;
Dirty: the cache line has been modified, but not yet written back to main memory;
Invalid: the data on the cache line is also present on other processors (it is not *reserved*), and another process has modified its copy of the data.

Exercise 1.9. Consider two processors, a data item x in memory, and cachelines x_1,x_2 in the private caches of the two processors to which x is mapped. Describe the transitions between the states of x_1 and x_2 under reads and writes of x on the two processors. Also indicate which actions cause memory bandwidth to be used. (This list of transitions is a Finite State Automaton (FSA); see section A.3.)

1.4 Locality and data reuse

By now it should be clear that there is more to the execution of an algorithm than counting the operations: the data transfer involved is important, and can in fact dominate the cost. Since we have caches and registers, the amount of data transfer can be minimized by programming in such a way that data stays as close to the processor as possible. Partly this is a matter of programming cleverly, but we can also look at the theoretical question: does the algorithm allow for it to begin with.

1.4.1 Data reuse

In this section we will take a look at *data reuse*: are data items involved in a calculation used more than once, so that caches and registers can be exploited? What precisely is the ratio between the number of operations and the amount of data transferred.

We define the data reuse ratio of an algorithm as follows:

> If n is the number of data items that an algorithm operates on, and $f(n)$ the number of operations it takes, then the reuse ratio is $f(n)/n$.

Consider for example the vector addition

$$\forall_i \colon x_i \leftarrow x_i + y_i.$$

This involves three memory accesses (two loads and one store) and one operation per iteration, giving a data reuse of $1/3$. The *axpy* (for '*a* times *x* plus *y*) operation

$$\forall_i \colon x_i \leftarrow x_i + a \cdot y_i$$

has two operations, but the same number of memory access since the one-time load of a is amortized. It is therefore more efficient than the simple addition, with a reuse of $2/3$.

The inner product calculation

$$\forall_i \colon s \leftarrow s + x_i \cdot y_i$$

is similar in structure to the axpy operation, involving one multiplication and addition per iteration, on two vectors and one scalar. However, now there are only two load operations, since s can be kept in register and only written back to memory at the end of the loop. The reuse here is 1.

Next, consider the matrix-matrix product:

$$\forall_{i,j} \colon c_{ij} = \sum_k a_{ik} b_{kj}.$$

This involves $3n^2$ data items and $2n^3$ operations, which is of a higher order. The data reuse is $O(n)$, meaning that every data item will be used $O(n)$ times. This has the implication that, with suitable programming, this operation has the potential of overcoming the bandwidth/clock speed gap by keeping data in fast cache memory.

Exercise 1.10. The matrix-matrix product, considered *as operation*, clearly has data reuse by the above definition. Argue that this reuse is not trivially attained by a simple implementation. What determines whether the naive implementation has reuse of data that is in cache?

[In this discussion we were only concerned with the number of operations of a given *implementation*, not the mathematical *operation*. For instance, there are ways or performing the matrix-matrix multiplication and Gaussian elimination algorithms in fewer than $O(n^3)$ operations [76, 69]. However, this requires a different implementation, which has its own analysis in terms of memory access and reuse.]

The matrix-matrix product is the heart of the 'LINPACK benchmark' [26]; see section 2.9. The benchmark may give an optimistic view of the performance of a computer: the matrix-matrix product is an operation that has considerable data reuse, so it is relatively insensitive to memory bandwidth and, for parallel computers, properties of the network. Typically, computers will attain 60–90% of their *peak performance* on the Linpack benchmark. Other benchmark may give considerably lower figures.

1.4.2 Locality

Since using data in cache is cheaper than getting data from main memory, a programmer obviously wants to code in such a way that data in cache is reused. While placing data in cache is not under explicit programmer control, even from assembly language, in most CPUs[7], it is still possible, knowing the behaviour of the caches, to know what data is in cache, and to some extent to control it.

The two crucial concepts here are *temporal locality* and *spatial locality*. Temporal locality is the easiest to explain: this describes the use of a data element within a short time of its last use. Since most caches have a LRU replacement policy (section 1.2.4.2), if in between the two references less data has been referenced than the cache size, the element will still be in cache and therefore be quickly accessible. With other replacement policies, such as random replacement, this guarantee can not be made.

As an example of temporal locality, consider the repeated use of a long vector:

7. Low level memory access can ben controlled by the programmer in the Cell processor and in some GPUs.

```
for (loop=0; loop<10; loop++) {
  for (i=0; i<N; i++) {
    ... = ... x[i] ...
  }
}
```

Each element of `x` will be used 10 times, but if the vector (plus other data accessed) exceeds the cache size, each element will be flushed before its next use. Therefore, the use of `x[i]` does not exhibit temporal locality: subsequent uses are spaced too far apart in time for it to remain in cache.

If the structure of the computation allows us to exchange the loops:

```
for (i=0; i<N; i++) {
  for (loop=0; loop<10; loop++) {
    ... = ... x[i] ...
  }
}
```

the elements of `x` are now repeatedly reused, and are therefore more likely to remain in the cache. This rearranged code displays better temporal locality in its use of `x[i]`.

The concept of *spatial locality* is slightly more involved. A program is said to exhibit spatial locality if it references memory that is 'close' to memory it already referenced. In the classical von Neumann architecture with only a processor and memory, spatial locality should be irrelevant, since one address in memory can be as quickly retrieved as any other. However, in a modern CPU with caches, the story is different. Above, you have seen two examples of spatial locality:

- Since data is moved in *cache line*s rather than individual words or bytes, there is a great benefit to coding in such a manner that all elements of the cacheline are used. In the loop
  ```
  for (i=0; i<N*s; i+=s) {
      ... x[i] ...
  }
  ```
  spatial locality is a decreasing function of the *stride* `s`.
  Let `S` be the cacheline size, then as `s` ranges from $1 \ldots$ `S`, the number of elements used of each cacheline goes down from `S` to 1. Relatively speaking, this increases the cost of memory traffic in the loop: if $s = 1$, we load $1/S$ cachelines per element; if $s = S$, we load one cacheline for each element. This effect is demonstrated in section 1.5.3.
- A second example of spatial locality worth observing involves the TLB (section 1.2.7). If a program references elements that are close together, they are likely on the same memory page, and address translation through the TLB will be fast. On the other hand, if a program references many widely disparate elements, it will also be referencing many different pages. The resulting TLB misses are very costly; see also section 1.5.4.

Let us examine locality issues for a realistic example. The matrix-matrix multiplicatio $C \leftarrow C + A \cdot B$ can be computed in several ways. We compare two implementations, assuming that all matrices are stored by rows, and

that the cache size is insufficient to store a whole row or column.

```
for i=1..n                          for i=1..n
  for j=1..n                          for k=1..n
    for k=1..n                          for j=1..n
      c[i,j] = c[i,j]                     c[i,j] = c[i,j]
               + a[i,k]*b[k,j]                     + a[i,k]*b[k,j]
```

Our first observation is that both implementations indeed compute $C \leftarrow C + A \cdot B$, and that they both take roughly $2n^3$ operations. However, their memory behaviour, including spatial and temporal locality is very different.

`c[i,j]` In the first implementation, `c[i,j]` is invariant in the inner iteration, which constitutes temporal locality, so it can be kept in register. As a result, each element of C will be loaded and stored only once.
In the second implementation, `c[i,j]` will be loaded and stored in each inner iteration. In particular, this implies that there are now n^3 store operations, a factor of n more than the first implementation.

`a[i,k]` In both implementations, `a[i,j]` elements are accessed by rows, so there is good spatial locality, as each loaded cacheline will be used entirely. In the second implementation, `a[i,k]` is invariant in the inner loop, which constitutes temporal locality; it can be kept in register. As a result, in the second case A will be loaded only once, as opposed to n times in the first case.

`b[k,j]` The two implementations differ greatly in how they access the matrix B. First of all, `b[k,j]` is never invariant so it will not be kept in register, and B engenders n^3 memory loads in both cases. However, the access patterns differ.
In second case, `b[k,j]` is access by rows so there is good spatial locality: cachelines will be fully utilized after they are loaded.
In the first implementation, `b[k,j]` is accessed by columns. Because of the row storage of the matrices, a cacheline contains a part of a row, so for each cacheline loaded, only one element is used in the columnwise traversal. This means that the first implementation has more loads for B by a factor of the cacheline length. There may also be TLB effects.

Note that we are not making any absolute predictions on code performance for these implementations, or even relative comparison of their runtimes. Such predictions are very hard to make. However, the above discussion identifies issues that are relevant for a wide range of classical CPUs.

Exercise 1.11. Consider the following pseudocode of an algorithm for summing n numbers $x[i]$ where n is a power of 2:

```
for s=2,4,8,...,n/2,n:
  for i=0 to n-1 with steps s:
    x[i] = x[i] + x[i+s/2]
sum = x[0]
```

Analyze the spatial and temporal locality of this algorithm, and contrast it with the standard algorithm

```
sum = 0
for i=0,1,2,...,n-1
  sum = sum + x[i]
```

1.5 Programming strategies for high performance

In this section we will look at how different ways of programming can influence the performance of a code. This will only be an introduction to the topic; for further discussion see the book by Goedeker and Hoisie [40].

The full listings of the codes and explanations of the data graphed here can be found in appendix D. All performance results were obtained on the AMD processors of the Ranger computer [9].

1.5.1 Pipelining

In section 1.1.1.1 you learned that the floating point units in a modern CPU are pipelined, and that pipelines require a number of independent operations to function efficiently. The typical pipelineable operation is a vector addition; an example of an operation that can not be pipelined is the inner product accumulation

```
for (i=0; i<N; i++)
  s += a[i]*b[i]
```

The fact that `s` gets both read and written halts the addition pipeline. One way to fill the *floating point pipeline* is to apply *loop unrolling*:

```
for (i = 0; i < N/2-1; i ++) {
  sum1 += a[2*i] * b[2*i];
  sum2 += a[2*i+1] * b[2*i+1];
}
```

Now there are two independent multiplies in between the accumulations. With a little indexing optimization this becomes:

```
for (i = 0; i < N/2-1; i ++) {
  sum1 += *(a + 0) * *(b + 0);
  sum2 += *(a + 1) * *(b + 1);

  a += 2; b += 2;
}
```

A first observation about this code is that we are implicitly using associativity and commutativity of addition: while the same quantities are added, they are now in effect added in a different order. As you will see in chapter 3, in computer arithmetic this is not guaranteed to give the exact same result.

In a further optimization, we disentangle the addition and multiplication part of each instruction. The hope is that while the accumulation is waiting for the result of the multiplication, the intervening instructions will keep the processor busy, in effect increasing the number of operations per second.

```
for (i = 0; i < N/2-1; i ++) {
  temp1 = *(a + 0) * *(b + 0);
  temp2 = *(a + 1) * *(b + 1);

  sum1 += temp1; sum2 += temp2;
```

```
    a += 2; b += 2;
  }
```

Finally, we realize that the furthest we can move the addition away from the multiplication, is to put it right in front of the multiplication *of the next iteration*:

```
  for (i = 0; i < N/2-1; i ++) {

    sum1 += temp1;
    temp1 = *(a + 0) * *(b + 0);

    sum2 += temp2;
    temp2 = *(a + 1) * *(b + 1);

    a += 2; b += 2;
  }

  s = temp1 + temp2;
```

Of course, we can unroll the operation by more than a factor of two. While we expect an increased performance because of the longer sequence of pipelined operations, large unroll factors need large numbers of registers. Asking for more registers than a CPU has is called *register spill*, and it will decrease performance.

Another thing to keep in mind is that the total number of operations is unlikely to be divisible by the unroll factor. This requires *cleanup code* after the loop to account for the final iterations. Thus, unrolled code is harder to write than straight code, and people have written tools to perform such *source-to-source transformations* automatically.

Cycle times for unrolling the inner product operation up to six times are given in table 1.2. Note that the timings do not show a monotone behaviour at the unrolling by four. This sort of variation is due to various memory-related factors.

1	2	3	4	5	6
6794	507	340	359	334	528

Table 1.2: Cycle times for the inner product operation, unrolled up to six times

1.5.2 Cache size

Above, you learned that data from L1 can be moved with lower latency and higher bandwidth than from L2, and L2 is again faster than L3 or memory. This is easy to demonstrate with code that repeatedly access the same data:

```
for (i=0; i<NRUNS; i++)
  for (j=0; j<size; j++)
    array[j] = 2.3*array[j]+1.2;
```

If the size parameter allows the array to fit in cache, the operation will be relatively fast. As the size of the dataset grows, parts of it will evict other parts from the L1 cache, so the speed of the operation will be determined by the latency and bandwidth of the L2 cache. This can be seen in figure 1.5. The full code is given in section D.2.

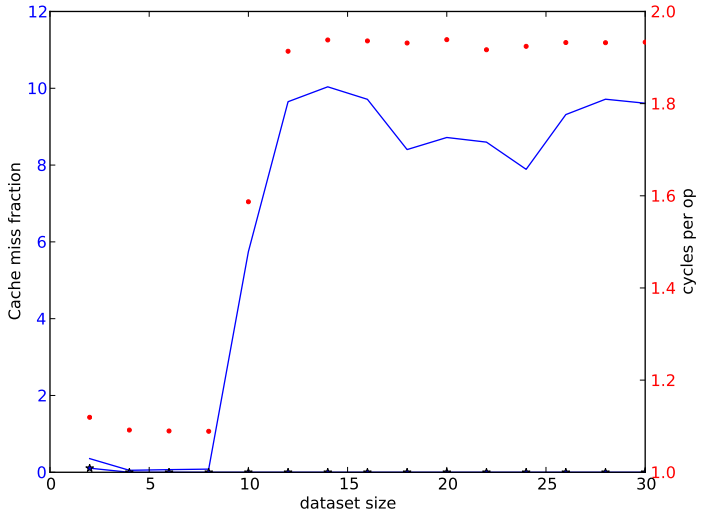


Figure 1.5: Average cycle count per operation as function of the dataset size

Exercise 1.12. Argue that with a large enough problem and an LRU replacement policy (section 1.2.4.2) essentially all data in the L1 will be replaced in every iteration of the outer loop. Can you write an example code that will let some of the L1 data stay resident?

Often, it is possible to arrange the operations to keep data in L1 cache. For instance, in our example, we could write

```
for (i=0; i<NRUNS; i++) {
  blockstart = 0;
  for (b=0; b<size/l1size; b++)
    for (j=0; j<l1size; j++)
      array[blockstart+j] = 2.3*array[blockstart+j]+1.2;
}
```

assuming that the L1 size divides evenly in the dataset size. This strategy is called *cache blocking* or *blocking for cache reuse*.

1.5.3 Cache lines

Since data is moved from memory to cache in consecutive chunks named cachelines (see section 1.2.4.3), code that does not utilize all data in a cacheline pays a bandwidth penalty. This is born out by a simple code

```
for (i=0,n=0; i<L1WORDS; i++,n+=stride)
  array[n] = 2.3*array[n]+1.2;
```

Here, a fixed number of operations is performed, but on elements that are at distance `stride`. As this *stride* increases, we expect an increasing runtime, which is born out by the graph in figure 1.6.

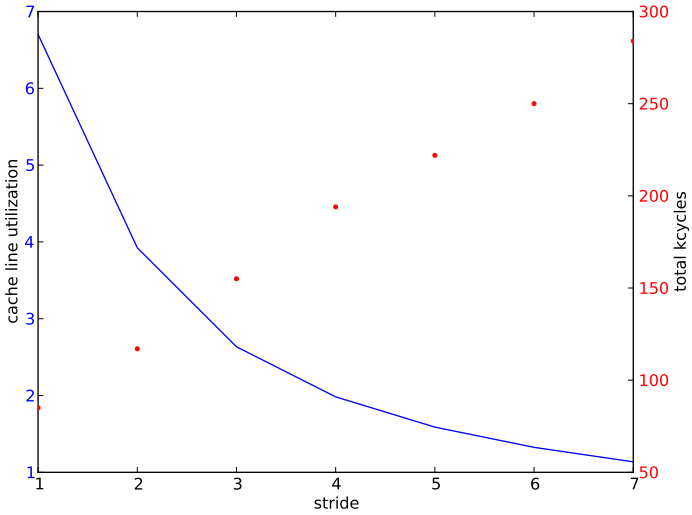


Figure 1.6: Run time in kcycles and L1 reuse as a function of stride

The graph also shows a decreasing reuse of cachelines, defined as the number of vector elements divided by the number of L1 misses (on stall; see section 1.2.5).

The full code is given in section D.3.

1.5.4 TLB

As explained in section 1.2.7, the Translation Look-aside Buffer (TLB) maintains a small list of frequently used memory pages and their locations; addressing data that are location on one of these pages is much faster than data that are not. Consequently, one wants to code in such a way that the number of pages accessed is kept low.

Consider code for traversing the elements of a two-dimensional array in two different ways.

```
#define INDEX(i,j,m,n) i+j*m
  array = (double*) malloc(m*n*sizeof(double));

  /* traversal #1 */
  for (j=0; j<n; j++)
```

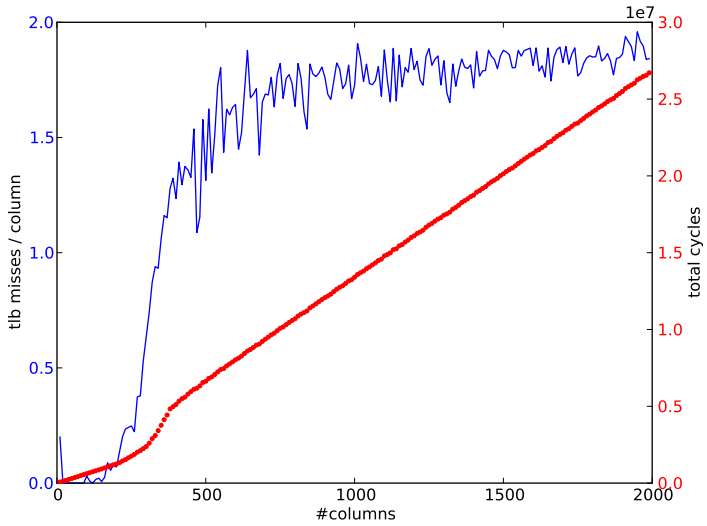


Figure 1.7: Number of TLB misses per column as function of the number of columns; columnwise traversal of the array.

```
    for (i=0; i<m; i++)
      array[INDEX(i,j,m,n)] = array[INDEX(i,j,m,n)]+1;

  /* traversal #2 */
  for (i=0; i<m; i++)
    for (j=0; j<n; j++)
      array[INDEX(i,j,m,n)] = array[INDEX(i,j,m,n)]+1;
```

The results (see Appendix D.5 for the source code) are plotted in figures 1.8 and 1.7.

Using $m = 1000$ means that, on the which has pages of 512 doubles, we need roughly two pages for each column. We run this example, plotting the number 'TLB misses', that is, the number of times a page is referenced that is not recorded in the TLB.

1. In the first traversal this is indeed what happens. After we touch an element, and the TLB records the page it is on, all other elements on that page are used subsequently, so no further TLB misses occur. Figure 1.8 shows that, with increasing n, the number of TLB misses per column is roughly two.
2. In the second traversal, we touch a new page for every element of the first row. Elements of the second row will be on these pages, so, as long as the number of columns is less than the number of TLB entries, these pages will still be recorded in the TLB. As the number of columns grows, the number of TLB increases, and ultimately there will be one TLB miss for each element access. Figure 1.7 shows that, with a large enough number of columns, the number of TLB misses per column is equal to the number of elements per column.

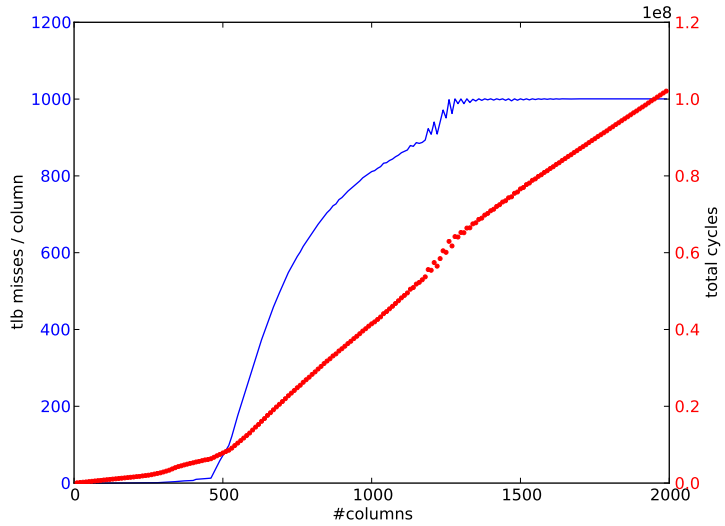


Figure 1.8: Number of TLB misses per column as function of the number of columns; rowwise traversal of the array.

1.5.5 Cache associativity

There are many algorithms that work by recursive division of a problem, for instance the *Fast Fourier Transform (FFT)* algorithm. As a result, code for such algorithms will often operate on vectors whose length is a power of two. Unfortunately, this can cause conflicts with certain architectural features of a CPU, many of which involve powers of two.

Consider the operation of adding a small number of vectors

$$\forall_j \colon y_j = y_j + \sum_{i=1}^{m} x_{i,j}.$$

If the length of the vectors y, x_i is precisely the right (or rather, wrong) number, y_j and $x_{i,j}$ will all be mapped to the same location in cache. As an example we take the AMD , which has an L1 cache of 64K bytes, and which is two-way set associative. Because of the set associativity, the cache can handle two addresses being mapped to the same cache location, but not three or more. Thus, we let the vectors be of size $n = 4096$ doubles, and we measure the effect in cache misses and cycles of letting $m = 1, 2, \ldots$.

First of all, we note that we use the vectors sequentially, so, with a cacheline of eight doubles, we should ideally see a cache miss rate of $1/8$ times the number of vectors m. Instead, in figure 1.9 we see a rate approximately proportional to m, meaning that indeed cache lines are evicted immediately. The exception here is the case $m = 1$, where the two-way associativity allows the cachelines of two vectors to stay in cache.

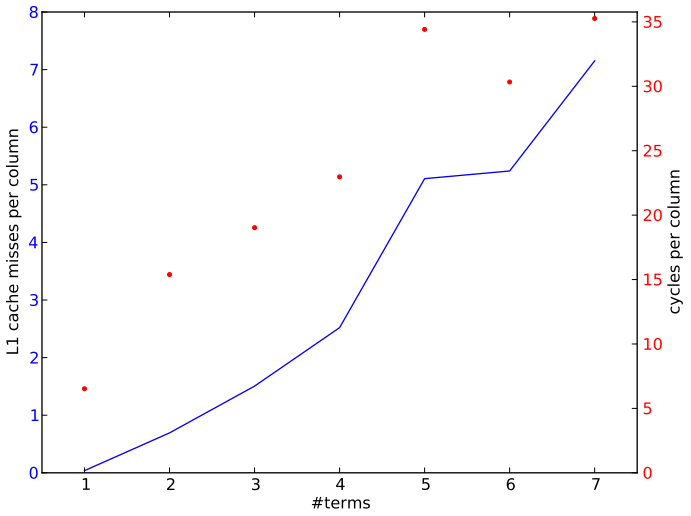


Figure 1.9: The number of L1 cache misses and the number of cycles for each j column accumulation, vector length 4096

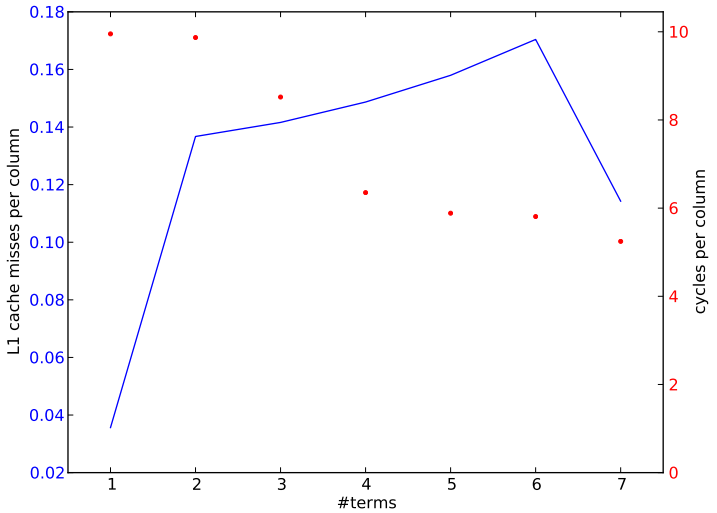


Figure 1.10: The number of L1 cache misses and the number of cycles for each j column accumulation, vector length $4096 + 8$

Compare this to figure 1.10, where we used a slightly longer vector length, so that locations with the same j are no longer mapped to the same cache location. As a result, we see a cache miss rate around $1/8$, and a smaller number of cycles, corresponding to a complete reuse of the cache lines.

Two remarks: the cache miss numbers are in fact lower than the theory predicts, since the processor will use prefetch streams. Secondly, in figure 1.10 we see a decreasing time with increasing m; this is probably due to a progressively more favourable balance between load and store operations. Store operations are more expensive than loads, for various reasons.

1.5.6 Loop tiling

```
for (n=0; n<10; n++)
  for (i=0; i<100000; i++)
    x[i] = ...

for (b=0; b<100; b++)
  for (n=0; n<10; n++)
    for (i=b*1000; i<(b+1)*1000; i++)
      x[i] = ...
```

1.5.7 Case study: Matrix-vector product

Let us consider in some detail the matrix-vector product

$$\forall_{i,j}: y_i \leftarrow a_{ij} \cdot x_j$$

This involves $2n^2$ operations on $n^2 + 2n$ data items, so reuse is $O(1)$: memory accesses and operations are of the same order. However, we note that there is a double loop involved, and the x, y vectors have only a single index, so each element in them is used multiple times.

Exploiting this theoretical reuse is not trivial. In

```
/* variant 1 */
for (i)
  for (j)
    y[i] = y[i] + a[i][j] * x[j];
```

the element `y[i]` seems to be reused. However, the statement as given here would write `y[i]` to memory in every inner iteration, and we have to write the loop as

```
/* variant 2 */
for (i) {
  s = 0;
  for (j)
    s = s + a[i][j] * x[j];
```

```
  y[i] = s;
}
```

to ensure reuse. This variant uses $2n^2$ loads and n stores.

This code fragment only exploits the reuse of `y` explicitly. If the cache is too small to hold the whole vector `x` plus a column of `a`, each element of `x` is still repeatedly loaded in every outer iteration.

Reversing the loops as

```
/* variant 3 */
for (j)
  for (i)
    y[i] = y[i] + a[i][j] * x[j];
```

exposes the reuse of `x`, especially if we write this as

```
/* variant 3 */
for (j) {
  t = x[j];
  for (i)
    y[i] = y[i] + a[i][j] * t;
}
```

but now `y` is no longer reused. Moreover, we now have $2n^2 + n$ loads, comparable to variant 2, but n^2 stores, which is of a higher order.

It is possible to get reuse both of x and y, but this requires more sophisticated programming. The key here is split the loops into blocks. For instance:

```
for (i=0; i<M; i+=2) {
  s1 = s2 = 0;
  for (j) {
    s1 = s1 + a[i][j] * x[j];
    s2 = s2 + a[i+1][j] * x[j];
  }
  y[i] = s1; y[i+1] = s2;
}
```

This is also called *loop unrolling*, *loop tiling*, or *strip mining*. The amount by which you unroll loops is determined by the number of available registers.

1.5.8 Optimization strategies

Figures 1.11 and 1.12 show that there can be wide discrepancy between the performance of naive implementations of an operation (sometimes called the 'reference implementation'), and optimized implementations. Unfortunately, optimized implementations are not simple to find. For one, since they rely on blocking, their loop nests are double the normal depth: the matrix-matrix multiplication becomes a six-deep loop. Then, the optimal block size is dependent on factors like the target architecture.

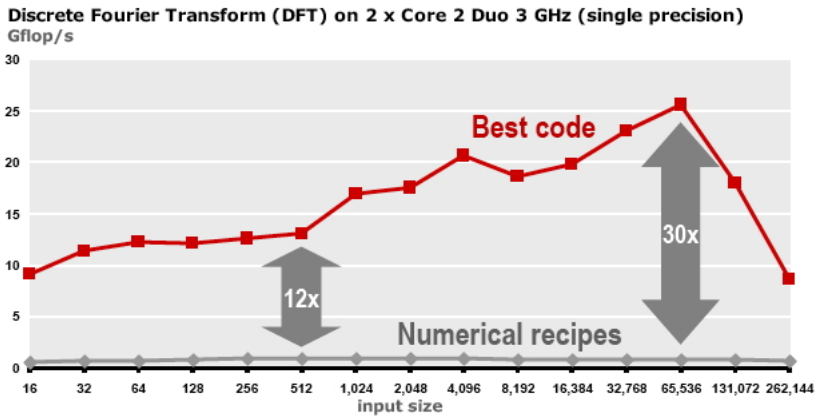


Figure 1.11: Performance of naive and optimized implementations of the Discrete Fourier Transform

We make the following observations:

- Compilers are not able to extract anywhere close to optimal performance[8].
- There are *autotuning* projects for automatic generation of implementations that are tuned to the architecture. This approach can be moderately to very successful. Some of the best known of these projects are Atlas [83] for Blas kernels, and Spiral [71] for transforms.

1.5.9 Cache aware programming

Unlike registers and main memory, both of which can be addressed in (assembly) code, use of caches is implicit. There is no way a programmer can load data explicitly to a certain cache, even in assembly language.

However, it is possible to code in a 'cache aware' manner. Suppose a piece of code repeatedly operates on an amount of data that less data than the cache size. We can assume that the first time the data is accessed, it is brought into cache; the next time it is accessed it will already be in cache. On the other hand, if the amount of data is more than the cache size[9], it will partly or fully be flushed out of cache in the process of accessing it.

We can experimentally demonstrate this phenomenon. With a very accurate counter, the code fragment

```
for (x=0; x<NX; x++)
  for (i=0; i<N; i++)
    a[i] = sqrt(a[i]);
```

will take time linear in `N` up to the point where `a` fills the cache. An easier way to picture this is to compute a normalized time, essentially a time per execution of the inner loop:

8. Presenting a compiler with the reference implementation may still lead to high performance, since some compilers are trained to recognize this operation. They will then forego translation and simply replace it by an optimized variant.
9. We are conveniently ignoring matters of set-associativity here, and basically assuming a fully associative cache.

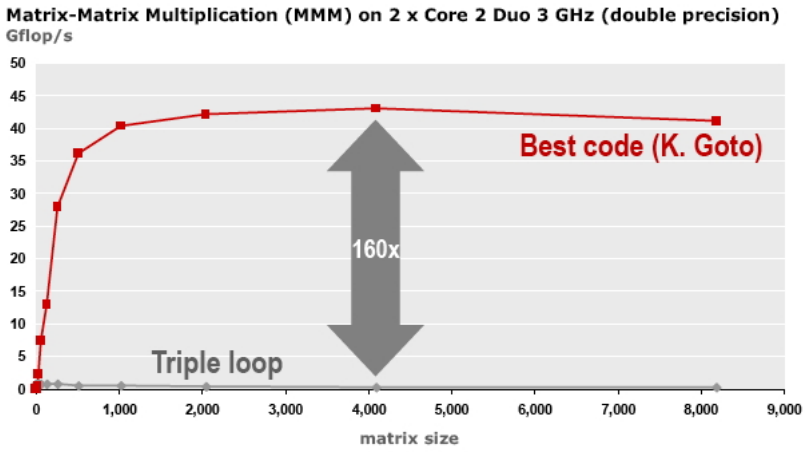


Figure 1.12: Performance of naive and optimized implementations of the matrix-matrix product

```
t = time();
for (x=0; x<NX; x++)
  for (i=0; i<N; i++)
    a[i] = sqrt(a[i]);
t = time()-t;
t_normalized = t/(N*NX);
```

The normalized time will be constant until the array `a` fills the cache, then increase and eventually level off again.

The explanation is that, as long as `a[0]...a[N-1]` fit in L1 cache, the inner loop will use data from the L1 cache. Speed of access is then determined by the latency and bandwidth of the L1 cache. As the amount of data grows beyond the L1 cache size, some or all of the data will be flushed from the L1, and performance will be determined by the characteristics of the L2 cache. Letting the amount of data grow even further, performance will again drop to a linear behaviour determined by the bandwidth from main memory.

Chapter 2

Parallel Computer Architecture

The largest and most powerful computers are sometimes called 'supercomputers'. For the last few decades, this has, without exception, referred to parallel computers: machines with more than one CPU that can be set to work on the same problem.

Parallelism is hard to define precisely, since it can appear on several levels. In the previous chapter you already saw how inside a CPU several instructions can be 'in flight' simultaneously. This is called *instruction-level parallelism*, and it is outside explicit user control: it derives from the compiler and the CPU deciding which instructions, out of a single instruction stream, can be processed simultaneously. At the other extreme is the sort of parallelism where more than one instruction stream is handled by multiple processors, often each on their own circuit board. This type of parallelism is typically explicitly scheduled by the user.

In this chapter, we will analyze this more explicit type of parallelism, the hardware that supports it, the programming that enables it, and the concepts that analyze it.

For further reading, a good introduction to parallel computers and parallel programming is Wilkinson and Allen [85].

2.1 Introduction

In scientific codes, there is often a large amount of work to be done, and it is often regular to some extent, with the same operation being performed on many data. The question is then whether this work can be sped up by use of a parallel computer. If there are n operations to be done, and they would take time t on a single processor, can they be done in time t/p on p processors?

Let us start with a very simple example. Adding two vectors of length n

```
for (i=0; i<n; i++)
  x[i] += y[i]
```

can be done with up to n processors, and execution time is linearly reduced with the number of processors. If each operation takes a unit time, the original algorithm takes time n, and the parallel execution on p processors n/p. The parallel algorithm is faster by a factor of p[1].

Next, let us consider summing the elements of a vector.

```
s = 0;
for (i=0; i<n; i++)
  s += x[i]
```

This is no longer obviously parallel, but if we recode the loop as

```
for (s=2; s<n; s*=2)
  for (i=0; i<n; i+=s)
    x[i] += x[i+s/2]
```

there is a way to parallelize it: every iteration of the outer loop is now a loop that can be done by n/s processors in parallel. Since the outer loop will go through $\log_2 n$ iterations, we see that the new algorithm has a reduced runtime of $n/p \cdot \log_2 n$. The parallel algorithm is now faster by a factor of $p/\log_2 n$.

Even from these two simple examples we can see some of the characteristics of parallel computing:

- Sometimes algorithms need to be rewritten slightly to make them parallel.
- A parallel algorithm may not show perfect speedup.

There are other things to remark on. In the first case, if there are $p = n$ processors, and each has its x_i, y_i in a local store, the algorithm can be executed without further complications. In the second case, processors need to *communicate* data among each other. What's more, if there is a concept of distance between processors, then each iteration of the outer loop increases the distance over which communication takes place.

These matters of algorithm adaptation, efficiency, and communication, are crucial to all of parallel computing. We will return to this issues in various guises throughout this chapter.

2.2 Parallel Computers Architectures

For quite a while now, the top computers have been some sort of parallel computer, that is, an architecture that allows the simultaneous execution of multiple instructions or instruction sequences. One way of characterizing the various forms this can take is due to Flynn [33]. Flynn's taxonomy distinguishes between whether one or more different instructions are executed simultaneously, and between whether that happens on one or more data items. The following four types result, which we will discuss in more detail below:

SISD Single Instruction Single Data: this is the traditional CPU architecture: at any one time only a single instruction is executed, operating on a single data item.

1. We ignore minor inaccuracies in this result when p does not divide perfectly in n. We will also, in general, ignore matters of loop overhead.

SIMD Single Instruction Multiple Data: in this computer type there can be multiple processors, each operating on its own data item, but they are all executing the same instruction on that data item. Vector computers (section 2.2.1.1) are typically also characterized as SIMD.

MISD Multiple Instruction Single Data. No architectures answering to this description exist.

MIMD Multiple Instruction Multiple Data: here multiple CPUs operate on multiple data items, each executing independent instructions. Most current parallel computers are of this type.

2.2.1 SIMD

Parallel computers of the SIMD type apply the same operation simultaneously to a number of data items. The design of the CPUs of such a computer can be quite simple, since the arithmetic unit does not need separate logic and instruction decoding units: all CPUs execute the same operation in lock step. This makes SIMD computers excel at operations on arrays, such as

```
for (i=0; i<N; i++) a[i] = b[i]+c[i];
```

and, for this reason, they are also often called *array processors*. Scientific codes can often be written so that a large fraction of the time is spent in array operations.

On the other hand, there are operations that can not can be executed efficiently on an array processor. For instance, evaluating a number of terms of a recurrence $x_{i+1} = ax_i + b_i$ involves that many additions and multiplications, but they alternate, so only one operation of each type can be processed at any one time. There are no arrays of numbers here that are simultaneously the input of an addition or multiplication.

In order to allow for different instruction streams on different parts of the data, the processor would have a 'mask bit' that could be set to prevent execution of instructions. In code, this typically looks like

```
where (x>0) {
  x[i] = sqrt(x[i])
```

The programming model where identical operations are applied to a number of data items simultaneously, is known as *data parallelism*.

Such array operations can occur in the context of physics simulations, but another important source is graphics applications. For this application, the processors in an array processor can be much weaker than the processor in a PC: often they are in fact bit processors, capable of operating on only a single bit at a time. Along these lines, ICL had the 4096 processor DAP [54] in the 1980s, and Goodyear built a 16K processor MPP [13] in the 1970s.

Later, the Connection Machine (CM-1, CM-2, CM-5) were quite popular. While the first Connection Machine had bit processors (16 to a chip), the later models had traditional processors capable of floating point arithmetic, and were not true SIMD architectures. All were based on a hyper-cube interconnection network; see section 2.6.4. Another manufacturer that had a commercially successful array processor was MasPar.

Supercomputers based on array processing do not exist anymore, but the notion of SIMD lives on in various guises, which we will now discuss.

2.2.1.1 Pipelining

A number of computers have been based on a *vector processor* or *pipeline processor* design. The first commercially successful supercomputers, the Cray-1 and the Cyber-205 were of this type. In recent times, the Cray-X1 and the NEC SX series have featured vector pipes. The 'Earth Simulator' computer [73], which led the TOP500 (section 2.9) for 3 years, was based on NEC SX processors. The general idea behind pipelining was described in section 1.1.1.1.

While supercomputers based on pipeline processors are in a distinct minority, pipelining is now mainstream in the superscalar CPUs that are the basis for clusters. A typical CPU has pipelined floating point units, often with separate units for addition and multiplication; see section 1.1.1.1.

However, there are some important differences between pipelining in a modern superscalar CPU and in, more old-fashioned, vector units. The pipeline units in these vector computers are not integrated floating point units in the CPU, but can better be considered as attached vector units to a CPU that itself has a floating point unit. The vector unit has vector registers[2] with a typical length of 64 floating point numbers; there is typically no 'vector cache'. The logic in vector units is also simpler, often addressable by explicit vector instructions. Superscalar CPUs, on the other hand, are more complicated and geared towards exploiting data streams in unstructured code.

2.2.1.2 True SIMD in CPUs and GPUs

True SIMD array processing can be found in modern CPUs and GPUs, in both cases inspired by the parallelism that is needed in graphics applications.

Modern CPUs from Intel and AMD, as well as PowerPC chips, have instructions that can perform multiple instances of an operation simultaneously. On Intel processors this is known as SSE: Streaming SIMD Extensions. These extensions were originally intended for graphics processing, where often the same operation needs to be performed on a large number of pixels. Often, the data has to be a total of, say, 128 bits, and this can be divided into two 64-bit reals, four 32-bit reals, or a larger number of even smaller chunks such as 4 bits.

Current compilers can generate SSE instructions automatically; sometimes it is also possible for the user to insert pragmas, for instance with the Intel compiler:

```
void func(float *restrict c, float *restrict a,
          float *restrict b, int n)
{
#pragma vector always
  for (int i=0; i<n; i++)
    c[i] = a[i] * b[i];
}
```

Use of these extensions often requires data to be aligned with cache line boundaries (section 1.2.4.3), so there are special allocate and free calls that return aligned memory.

2. The Cyber205 was an exception, with direct-to-memory architecture.

Array processing on a larger scale can be found in *GPU*s. A GPU contains a large number of simple processors, ordered in groups of 32, typically. Each processor group is limited to executing the same instruction. Thus, this is true example of Single Instruction Multiple Data (SIMD) processing.

2.2.2 MIMD / SPMD computers

By far the most common parallel computer architecture these days is called Multiple Instruction Multiple Data (MIMD): the processors execute multiple, possibly differing instructions, each on their own data. Saying that the instructions differ does not mean that the processors actually run different programs: most of these machines operate in Single Program Multiple Data (SPMD) mode, where the programmer starts up the same executable on the parallel processors. Since the different instances of the executable can take differing paths through conditional statements, or execute differing numbers of iterations of loops, they will in general not be completely in sync as they were on SIMD machines. If this lack of synchronization is due to processors working on different amounts of data, it is called *load unbalance*, and it is a major source of less than perfect *speedup*.

There is a great variety in MIMD computers. Some of the aspects concern the way memory is organized, and the network that connects the processors. Apart from these hardware aspects, there are also differing ways of programming these machines. We will see all these aspects below. Machines supporting the SPMD model are usually called *clusters*. They can be built out of custom or commodity processors; if they consist of PCs, running Linux, and connected with Ethernet, they are referred to as Beowulf clusters [47].

2.3 Different types of memory access

Most processors are considerably faster than the connections they have to memory. For parallel machines, where potentially several processors want to access the same memory location, this problem becomes even worse. We can characterize parallel machines by the approach they take to the problem of reconciling multiple accesses, by multiple processes, to a joint pool of data.

2.3.1 Symmetric Multi-Processors: Uniform Memory Access

Parallel programming is fairly simple if any processor can access any memory location. For this reason, there is a strong incentive for manufacturers to make architectures where processors see no difference between one memory location and another: they are all accessible, and the access times do not differ. This is called *Uniform Memory Access (UMA)*, and an architecture on this principle is often called an *Symmetric Multi Processing (SMP)*.

There are a few ways to realize an SMP architecture. Current desktop computers can have a few processors accessing a shared memory through a single memory bus; for instance Apple markets a model with 2 six-core processors. Having a memory bus that is shared between processors works only for small numbers of processors; for larger numbers one can use a *crossbar* that connects multiple processors to multiple memory banks; see section 2.6.5.

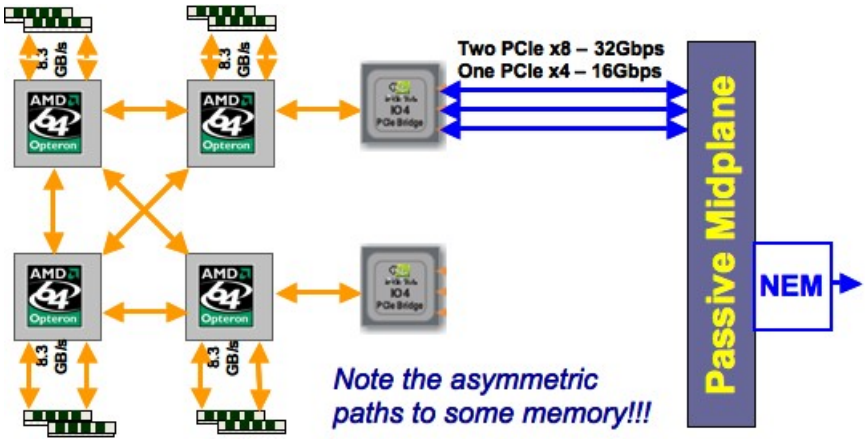


Figure 2.1: Non-uniform memory access in a four-socket motherboard

2.3.2 Non-Uniform Memory Access

The UMA approach based on shared memory is obviously limited to a small number of processors. The crossbar networks are expandable, so they would seem the best choice.

However, beyond a very small number of processing elements, UMA does not really apply since memoryless processors do not exist; in practice one puts processors with a local memory in a configuration with an exchange network. This leads to a situation where a processor can access its own memory, fast, and other processors' memory, slower. This is one case of so-called *Non-Uniform Memory Access (NUMA)*: a strategy that uses physically distributed memory, abandoning the uniform access time, but maintaining the logically shared memory: each processor can still access any memory location.

Figure 2.1 illustrates NUMA in the case of the four-socket motherboard of the Ranger supercomputer. Each chip has its own memory (8Gb) but the motherboard acts as if the processors have access to a shared pool of 32Gb. Obviously, accessing the memory of another processor is slower than accessing local memory. In addition, note that each processor has three connections that could be used to access other memory, but the rightmost two chips use one connection to connect to the network. This means that accessing each other's memory can only happen through an intermediate processor, slowing down the transfer, and tieing up that processor's connections.

While the NUMA approach is convenient for the programmer, it offers some challenges for the system. Imagine that two different processors each have a copy of a memory location in their local (cache) memory. If one processor alters the content of this location, this change has to be propagated to the other processors. If both processors try to alter the content of the one memory location, the behaviour of the program can become undetermined.

Keeping copies of a memory location synchronized is known as *cache coherence* (see section 1.3.1, and a multiprocessor system using it is sometimes called a 'cache-coherent NUMA' or *ccNUMA* architecture.

Cache coherence is obviously desirable, since it facilitates parallel programming. However, it carries costs on several levels. It probably requires hardware support from the processors and the network and it complicates the Operating

System software. Moreover, keeping caches coherent means that there is data traveling through the network, taking up precious bandwidth.

The SGI Origin and Onyx computers can have more than a 1000 processors in a NUMA architecture. This requires a substantial network.

2.3.3 Logically and physically distributed memory

The most extreme solution to the memory access problem is to offer memory that is not just physically, but that is also logically distributed: the processors have their own address space, and can not directly see another processor's memory. This approach is often called 'distributed memory', but note that NUMA also has physically distributed memory, the distributed nature of it is just not apparent to the programmer.

With logically distributed memory, the only way one processor can exchange information with another is through passing information explicitly through the network. You will see more about this in section 2.5.2.

This type of architecture has the significant advantage that it can scale up to large numbers of processors: the IBM Blue Gene has been built with over 200,000 processors. On the other hand, this is also the hardest kind of parallel system to program.

Various kinds of hybrids between the above types exist. For instance, the Columbia computer at NASA consists of twenty nodes, connected by a switch, where each node is an NUMA architecture of 512 processors.

2.4 Granularity of parallelism

Let us take a look at the question 'how much parallelism is there in a program execution'. There is the theoretical question of the absolutely maximum number of actions that can be taken in parallel, but we also need to wonder what kind of actions these are and how hard it is to actually execute them in parallel, as well has how efficient the resulting execution is.

The discussion in this section will be mostly on a conceptual level; in section 2.5 we will go into some detail on how parallelism can actually be programmed.

2.4.1 Data parallelism

It is fairly common for a program that have loops with a simple body, that gets executed for all elements in a large data set:

```
for (i=0; i<1000000; i++)
  a[i] = 2*b[i];
```

Such code is considered an instance of *data parallelism* or *fine-grained parallelism*. If you had as many processors as array elements, this code would look very simple: each processor would execute the statment

```
a = 2*b
```

on its local data.

If your code consists predominantly of such loops over arrays, it can be executed efficiently with all processors in lockstep. Architectures based on this idea, where the processors can in fact *only* work in lockstep, have existed, see section 2.2.1. Such fully parallel operations on arrays appear in computer graphics, where every bit of an image is processed independently. For this reason, GPUs are strongly based on data parallelism.

Continuing the above example for a little bit, consider the operation

for $0 \leq i < 10^6$ **do**
  $a_i = (b_{i-1} + b_{i+1})/2$
end

On a data parallel machine, that could be implemented as

`bleft` $\leftarrow$ **shiftright**(`b`)
`bright` $\leftarrow$ **shiftleft**(`b`)
`a` $\leftarrow$ (`bleft` + `bright`)$/2$

where the `shiftleft/right` instructions cause a data item to be sent to the processor with a number lower or higher by 1. For this second example to be efficient, it is necessary that each processor can communicate quickly with its immediate neighbours, and the first and last processor with each other.

In various contexts such a 'blurr' operations in graphics, it makes sense to have operations on 2D data:

for $0 < i < m$ **do**
  for $0 < j < n$ **do**
    $a_{ij} \leftarrow (b_{ij-1} + b_{ij+1} + b_{i-1j} + b_{i+1j})$
  end
end

and consequently processors have be able to move data to neighbours in a 2D grid.

2.4.2 Instruction-level parallelism

In *Instruction Level Parallelism (ILP)*, the parallelism is still on the level of individual instructions, but these need not be similar. For instance, in

$a \leftarrow b + c$
$d \leftarrow e + f$

the two assignments are independent, and can therefore be executed simultaneously. This kind of parallelism is too cumbersome for humans to identify, but compilers are very good at this. In fact, identifying ILP is crucial for getting good performance out of modern *superscalar* CPUs.

2.4.3 Task-level parallelism

At the other extreme from data and instruction-level parallelism, *task parallelism* is about identifying whole subprograms that can be executed in parallel. As an example, searching in a tree data structure could be implemented as follows:

```
if optimal (root) then
    exit
else
    parallel: SearchInTree (leftchild),SearchInTree (rightchild)
end
```

Procedure SearchInTree(root)

The search tasks in this example are not synchronized, and the number of tasks is not fixed: it can grow arbitrarily. In practice, having too many tasks is not a good idea, since processors are most efficient if they work on just a single task. Tasks can then be scheduled as follows:

```
while there are tasks left do
    wait until a processor becomes inactive;
    spawn a new task on it
end
```

Unlike in the data parallel example above, the assignment of data to processor is not determined in advance in such a scheme. Therefore, this mode of parallelism is most suited for thread-programming, for instance through the OpenMP library; section 2.5.1.

Let us consider a more serious example of task-level parallelism.

A finite element mesh is, in the simplest case, a collection of triangles that covers a 2D object. Since angles that are too acute should be avoided, the *Delauney mesh refinement* process can take certain triangles, and replace them by better shaped ones. This is illustrated in figure 2.2: the black triangles violate some angle condition, so either they themselves get subdivided, or they are joined with some neighbouring ones (rendered in grey) and then jointly redivided.

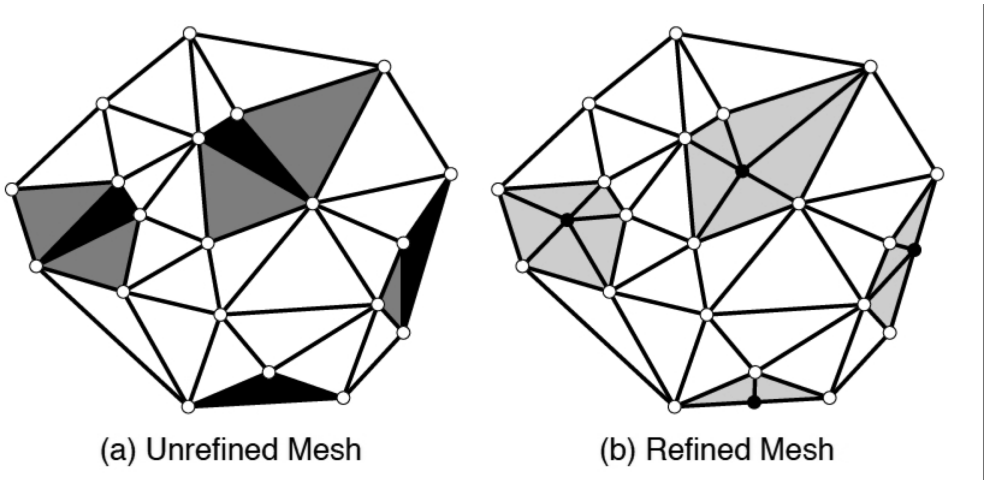


Figure 2.2: A mesh before and after refinement

In pseudo-code, this can be implemented as follows (for a more detailed discussion, see [57]):

```
Mesh m = /* read in initial mesh */
WorkList wl;
wl.add(mesh.badTriangles());
while (wl.size() != 0) do
    Element e = wl.get(); //get bad triangle
    if (e no longer in mesh) continue;
    Cavity c = new Cavity(e);
    c.expand();
    c.retriangulate();
    mesh.update(c);
    wl.add(c.badTriangles());
end
```

It is clear that this algorithm is driven by a worklist data structure that has to be shared between all processes. Together with the dynamic assignment of data to processes, this implies that this type of *irregular parallelism* is suited to shared memory programming, and is much harder to do with distributed memory.

2.4.4 Parameter sweeps

In certain contexts, a simple, often single processor, calculation needs to be performed on many different inputs. Scheduling the calculations as indicated above is then referred to as a *parameter sweep*. Since the program executions have no data dependencies and need not done in any particular sequence, this is an example of *embarassingly parallel* computing.

2.4.5 Medium-grain data parallelism

The above strict realization of data parallelism assumes that there are as many processors as data elements. In practice, processors will have much more memory than that, and the number of data elements is likely to be far larger than the processor count of even the largest computers. Therefore, arrays are grouped onto processors in subarrays. The code then looks like this:

```
my_lower_bound = // some processor-dependent number
my_upper_bound = // some processor-dependent number
for (i=my_lower_bound; i<my_upper_bound; i++)
  // the loop body goes here
```

This model has some characteristics of data parallelism, since the operation performed is identical on a large number of data items. It can also be viewed as task parallelism, since each processor executes a larger section of code, and does not necessarily operate on equal sized chunks of data.

2.5 Parallel programming

Parallel programming is more complicated than sequential programming. While for sequential programming most programming languages operate on similar principles (some exceptions such as functional or logic languages aside), there is a variety of ways of tackling parallelism. Let's explore some of the concepts and practical aspects.

There are various approaches to parallel programming. One might say that they fall in two categories:

- Let the programmer write a program that is essentially the same as a sequential program, and let the lower software and hardware layers solve the parallelism problem; and
- Expose the parallelism to the programmer and let the programmer manage everything explicitly.

The first approach, which is more user-friendly, is in effect only possible with shared memory, where all processes have access to the same address space. We will discuss this in the section on OpenMP 2.5.1. The second approach is necessary in the case of distributed memory programming. We will have a general discussion of distributed programming in section 2.5.2.1; section 2.5.2 will discuss the MPI library.

2.5.1 OpenMP

OpenMP is an extension to the programming languages C and Fortran. Its main approach to parallelism is the parallel execution of loops: based on compiler directives, a preprocessor can schedule the parallel execution of the loop iterations.

The amount of parallelism is flexible: the user merely specifies a parallel region, indicating that all iterations are independent to some extent, and the runtime system will then use whatever resources are available. Because of this dynamic nature, and because no data distribution is specified, OpenMP can only work with threads on shared memory.

OpenMP is neither a language nor a library: it operates by inserting directives into source code, which are interpreted by the compiler. Many compilers, such as GCC or the Intel compiler, support the OpenMP extensions. In Fortran,

OpenMP directives are placed in comment statements; in C, they are placed in `#pragma` CPP directives, which indicate compiler specific extensions. As a result, OpenMP code still looks like legal C or Fortran to a compiler that does not support OpenMP. Programs need to be linked to an OpenMP runtime library, and their behaviour can be controlled through environment variables.

OpenMP features *dynamic parallelism*: the number of execution streams operating in parallel can vary from one part of the code to another.

For more information about OpenMP, see [19].

2.5.1.1 Processes and threads

OpenMP is based on 'threads' working in parallel; see section 2.5.1.3 for an illustrative example. A *thread* is an independent instruction stream, but as part of a Unix process. While processes can belong to different users, or be different programs that a single user is running concurrently, and therefore have their own data space, threads are part of one process and therefore share each other's data. Threads do have a possibility of having private data, for instance, they have their own data stack.

Threads serve two functions:

1. By having more than one thread on a single processor, a higher processor utilization can result, since the instructions of one thread can be processed while another thread is waiting for data. On traditional CPUs, switching between threads is fairly expensive (an exception is the *hyper-threading* mechanism) but on GPUs it is not, and in fact they *need* many threads to attain high performance.
2. In a shared memory context, multiple threads running on multiple processors or processor cores can be an easy way to parallelize a process. The shared memory allows the threads to all see the same data.

2.5.1.2 Issues in shared memory programming

Shared memory makes life easy for the programmer, since every processor has access to all of the data: no explicit data traffic between the processor is needed. On the other hand, multiple processes/processors can also write to the same variable, which is a source of potential problems.

Suppose that two processes both try to increment an integer variable `I` by one:

process 1: `I=I+2`
process 2: `I=I+3`

If the processes are not completely synchronized, one will read the current value, compute the new value, write it back, and leave that value for the other processor to find. In this scenario, the parallel program has the same result (`I=I+5`) as if all instructions were executed sequentially.

However, it could also happen that both processes manage to read the current value simultaneously, compute their own new value, and write that back to the location of `I`. Even if the conflicting writes can be reconciled, the final result will be wrong: the new value will be either `I+2` or `I+3`, not `I+5`. Moreover, it will be indeterminate, depending on details of the execution mechanism.

For this reason, such updates of a shared variable are called a *critical section* of code. OpenMP has a mechanism to declare a critical section, so that it will be executed by only one process at a time. One way of implementing this, is to set a temporary *lock* on certain memory areas. Another solution to the update problem, is to have *atomic operations*: the update would be implemented in such a way that a second process can not get hold of the data item being updated. One implementation of this is *transactional memory*, where the hardware itself supports atomic operations; the term derives from database transactions, which have a similar integrity problem.

Finally, we mention the *semaphore* mechanism for dealing with critical sections [23]. Surrounding each critical section there will be two atomic operations controlling a semaphore. The first process to encounter the semaphore will lower it, and start executing the critical section. Other processes see the lowered semaphore, and wait. When the first process finishes the critical section, it executes the second instruction which raises the semaphore, allowing one of the waiting processes to enter the critical section.

2.5.1.3 Threads example

The following example spawns a number of tasks that all update a global counter. Since threads share the same memory space, they indeed see and update the same memory location.

```
#include <stdlib.h>
#include <stdio.h>
#include "pthread.h"

int sum=0;

void adder() {
  sum = sum+1;
  return;
}

#define NTHREADS 50
int main() {
  int i;
  pthread_t threads[NTHREADS];
  printf("forking\n");
  for (i=0; i<NTHREADS; i++)
    if (pthread_create(threads+i,NULL,&adder,NULL)!=0) return i+1;
  printf("joining\n");
  for (i=0; i<NTHREADS; i++)
    if (pthread_join(threads[i],NULL)!=0) return NTHREADS+i+1;
  printf("Sum computed: %d\n",sum);

  return 0;
```

```
}
```

The fact that this code gives the right result is a coincidence: it only happens because updating the variable is so much quicker than creating the thread. (On a multicore processor the chance of errors will greatly increase.) If we artificially increase the time for the update, we will no longer get the right result:

```
void adder() {
  int t = sum; sleep(1); sum = t+1;
  return;
}
```

Now all threads read out the value of `sum`, wait a while (presumably calculating something) and then update.

This can be fixed by having a lock on the code region that should be 'mutually exclusive':

```
pthread_mutex_t lock;

void adder() {
  int t,r;
  pthread_mutex_lock(&lock);
  t = sum; sleep(1); sum = t+1;
  pthread_mutex_unlock(&lock);
  return;
}

int main() {
  ....
  pthread_mutex_init(&lock,NULL);
```

The lock and unlock commands guarantee that no two threads can interfere with each other's update.

2.5.1.4 OpenMP examples

The simplest example of OpenMP use is the parallel loop.

```
  #pragma omp for
  for (i=0; i<ProblemSize; i++) {
    a[i] = b[i];
  }
```

Clearly, all iterations can be executed independently and in any order. The pragma CPP directive then conveys this fact to the compiler.

Some loops are fully parallel conceptually, but not in implementation:

```
  for (i=0; i<ProblemSize; i++) {
    t = b[i]*b[i];
```

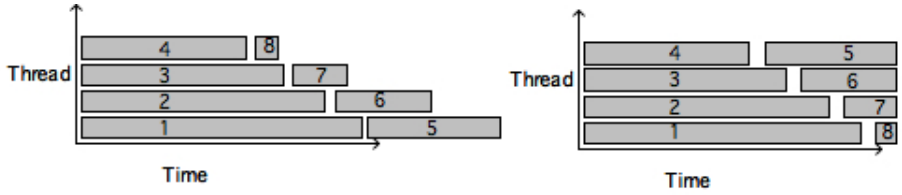


Figure 2.3: Static or round-robin (left) vs dynamic (right) thread scheduling; the task numbers are indicated.

```
    a[i] = sin(t) + cos(t);
  }
```

Here it looks as if each iteration writes to, and reads from, a shared variable `t`. However, `t` is really a temporary variable, local to each iteration. OpenMP indicates that as follows:

```
#pragma parallel for shared(a,b), private(t)
  for (i=0; i<ProblemSize; i++) {
    t = b[i]*b[i];
    a[i] = sin(t) + cos(t);
  }
```

If a scalar *is* indeed shared, OpenMP has various mechanisms for dealing with that. For instance, shared variables commonly occur in *reduction operations*:

```
  s = 0;
#pragma parallel for reduction(+:sum)
  for (i=0; i<ProblemSize; i++) {
    s = s + a[i]*b[i];
  }
```

As you see, a sequential code can be easily parallelized this way.

The assignment of iterations to threads is done by the runtime system, but the user can guide this assignment. We are mostly concerned with the case where there are more iterations than threads: if there are P threads and N iterations and $N > P$, how is iteration i going to be assigned to a thread?

The simplest assignment uses *round-robin scheduling*, a *static scheduling* strategy where thread p get iterations $p, p + P, p + 2P, \ldots$. This has the advantage that if some data is reused between iterations, it will stay in the data cache of the processor executing that thread. On the other hand, if the iterations differ in the amount of work involved, the process may suffer from *load unbalance* with static scheduling. In that case, a *dynamic scheduling* strategy would work better, where each thread starts work on the next unprocessed iteration as soon as it finishes its current iteration. The example in figure 2.3 shows static versus dynamic scheduling of a number of tasks that gradually decrease in individual running time. In static scheduling, the first thread gets tasks 1 and 4, the second 2 and 5, et cetera. In dynamic scheduling, any thread that finishes its task gets the next task. This clearly gives a better running time in this particular example.

2.5.2 Message passing through MPI

While OpenMP programs, and programs written using other shared memory paradigms, still look very much like sequential programs, this does not hold true for message passing code. Before we discuss the Message Passing Interface (MPI) library in some detail, we will take a look at this shift the way parallel code is written.

2.5.2.1 The global versus the local view in distributed programming

There can be a marked difference between how a parallel algorithm looks to an observer, and how it is actually programmed. Consider the case where we have an array of processors $\{P_i\}_{i=0..p-1}$, each containing one element of the arrays x and y, and P_i computes

$$\begin{cases} y_i \leftarrow y_i + x_{i-1} & i > 0 \\ y_i \text{ unchanged} & i = 0 \end{cases} \tag{2.1}$$

The global description of this could be

- Every processor P_i except the last sends its x element to P_{i+1};
- every processor P_i except the first receive an x element from their neighbour P_{i-1}, and
- they add it to their y element.

However, in general we can not code in these global terms. In the SPMD model (section 2.2.2) each processor executes the same code, and the overall algorithm is the result of these individual behaviours. The local program has access only to local data – everything else needs to be communicated with send and receive operations – and the processor knows its own number.

A naive attempt at the processor code would look like:

- If I am processor 0, do nothing; otherwise
- receive an x element from my left neighbour,
- add it to my y element, and
- send my x element to my right neighbour, unless I am the last processor.

This code is correct, but it is also inefficient: processor $i + 1$ does not start working until processor i is done. In other words, the parallel algorithm is in effect sequential, and offers no speedup.

We can easily make our algorithm fully parallel, by changing the processor code to:

- If am not the last processor, send my x element to the right.
- If am not the first processor, receive an x element from the left, and
- add it to my y element.

Now all sends, receives, and additions can happen in parallel.

There can still be a problem with this solution if the sends and receives are so-called *blocking communication* instructions: a send instruction does not finish until the sent item is actually received, and a receive instruction waits for the corresponding send. In this scenario:

- All processors but the last start their send instruction, and consequently block, waiting for their neighbour to execute the corresponding receive.
- The last processor is only one not sending, so it executes its receive instruction.
- The processor one-before-last now completes its send, and progresses to its receive;
- Which allows its left neighbour to complete its send, et cetera.

You see that again the parallel execution becomes serialized.

Exercise 2.1. Suggest a way to make the algorithm parallel, even with blocking operations. (Hint: you won't be able to get all processors active at the same time.)

If the algorithm in equation 2.1 had been cyclic:

$$\begin{cases} y_i \leftarrow y_i + x_{i-1} & i = 1 \ldots n-1 \\ y_0 \leftarrow y_0 + x_{n-1} & i = 0 \end{cases}$$

the problem would be even worse. Now the last processor can not start its receive since it is blocked sending x_{n-1} to processor 0. This situation, where the program can not progress because every processor is waiting for another, is called *deadlock*.

The reason for blocking instructions is to prevent accumulation of data in the network. If a send instruction were to complete before the corresponding receive started, the network would have to store the data somewhere in the mean time. Consider a simple example:

```
buffer = ... ;  // generate some data
send(buffer,0); // send to processor 0
buffer = ... ;  // generate more data
send(buffer,1); // send to processor 1
```

After the first send, we start overwriting the buffer. If the data in it hasn't been received, the first set of values would have to be buffered somewhere in the network, which is not realistic. By having the send operation block, the data stays in the sender's buffer until it is guaranteed to have been copied to the recipient's buffer.

One way out of the problem of sequentialization or deadlock that arises from blocking instruction is the use of *non-blocking communication* instructions, which include explicit buffers for the data. With non-blocking send instruction, the user needs to allocate a buffer for each send, and check when it is safe to overwrite the buffer.

```
buffer0 = ... ;   // data for processor 0
send(buffer0,0);  // send to processor 0
buffer1 = ... ;   // data for processor 1
send(buffer1,1);  // send to processor 1
...
// wait for completion of all send operations.
```

2.5.2.2 The MPI library

If OpenMP is the way to program shared memory, *MPI* [75] is the standard solution for programming distributed memory. MPI ('Message Passing Interface') is a specification for a library interface for moving between processes

that do not otherwise share data. The MPI routines can be divided roughly in the following categories:

- Process management. This includes querying the parallel environment and constructing subsets of processors.
- Point-to-point communication. This is a set of calls where two processes interact. These are mostly variants of the send and receive calls.
- Collective calls. In these routines, all processors (or the whole of a specified subset) are involved. Examples are the *broadcast* call, where one processor shares its data with every other processor, or the *gather* call, where one processor collects data from all participating processors.

Let us consider how the OpenMP examples can be coded in MPI. First of all, we no longer allocate

```
double a[ProblemSize];
```

but

```
double a[LocalProblemSize];
```

where the local size is roughly a $1/P$ fraction of the global size. (Practical considerations dictate whether you want this distribution to be as evenly as possible, or rather biased in some way.)

The parallel loop is trivially parallel, with the only difference that it now operates on a fraction of the arrays:

```
for (i=0; i<LocalProblemSize; i++) {
  a[i] = b[i];
}
```

However, if the loop involves a calculation based on the iteration number, we need to map that to the global value:

```
for (i=0; i<LocalProblemSize; i++) {
  a[i] = b[i]+f(i+MyFirstVariable);
}
```

(We will assume that each process has somehow calculated the values of `LocalProblemSize` and `MyFirstVariable`.) Local variables are now automatically local, because each process has its own instance:

```
for (i=0; i<LocalProblemSize; i++) {
  t = b[i]*b[i];
  a[i] = sin(t) + cos(t);
}
```

However, shared variables are harder to implement. Since each process has its own data, the local accumulation has to be explicitly assembled:

```
for (i=0; i<LocalProblemSize; i++) {
  s = s + a[i]*b[i];
}
MPI_Allreduce(s,globals,1,MPI_DOUBLE,MPI_SUM);
```

The 'reduce' operation sums together all local values `s` into a variable `globals` that receives an identical value on each processor. This is known as a *collective operation*.

Let us make the example slightly more complicated:

```
for (i=0; i<ProblemSize; i++) {
  if (i==0)
    a[i] = (b[i]+b[i+1])/2
  else if (i==ProblemSize-1)
    a[i] = (b[i]+b[i-1])/2
  else
    a[i] = (b[i]+b[i-1]+b[i+1])/3
```

The basic form of the parallel loop is:

```
for (i=0; i<LocalProblemSize; i++) {
  bleft = b[i-1]; bright = b[i+1];
  a[i] = (b[i]+bleft+bright)/3
```

First we account for the fact that `bleft` and `bright` need to be obtained from a different processor for `i==0` (`bleft`), and for `i==LocalProblemSize-1` (`bright`). We do this with a exchange operation with our left and right neighbour processor:

```
// get bfromleft and bfromright from neighbour processors, then
for (i=0; i<LocalProblemSize; i++) {
  if (i==0) bleft=bfromleft;
    else bleft = b[i-1]
  if (i==LocalProblemSize-1) bright=bfromright;
    else bright = b[i+1];
  a[i] = (b[i]+bleft+bright)/3
```

Obtaining the neighbour values is done as follows. First we need to ask our processor number, so that we can start a communication with the processor with a number one higher and lower.

```
MPI_Comm_rank(MPI_COMM_WORLD,&myTaskID);
MPI_Sendrecv
  (/* to be sent: */ &b[LocalProblemSize-1],
   /* result:     */ &bfromleft,
   /* destination */ myTaskID+1, /* some parameters omited */ );
MPI_Sendrecv(&b[0],&bfromright,myTaskID-1 /* ... */ );
```

There are still two problems with this code. First, the sendrecv operations need exceptions for the first and last processors. This can be done elegantly as follows:

```
MPI_Comm_rank(MPI_COMM_WORLD,&myTaskID);
MPI_Comm_size(MPI_COMM_WORLD,&nTasks);
if (myTaskID==0) leftproc = MPI_PROC_NULL;
  else leftproc = myTaskID-1;
if (myTaskID==nTasks-1) rightproc = MPI_PROC_NULL;
  else rightproc = myTaskID+1;
MPI_Sendrecv( &b[LocalProblemSize-1], &bfromleft,  rightproc );
MPI_Sendrecv( &b[0],                  &bfromright, leftproc);
```

Exercise 2.2. There is still a problem left with this code: the boundary conditions from the original, global, version have not been taken into account. Give code that solves that problem.

MPI gets complicated if different processes need to take different actions, for example, if one needs to send data to another. The problem here is that each process executes the same executable, so it needs to contain both the send and the receive instruction, to be executed depending on what the rank of the process is.

```
if (myTaskID==0) {
  MPI_Send(myInfo,1,MPI_INT,/* to: */ 1,/* labeled: */,0,
    MPI_COMM_WORLD);
} else {
  MPI_Recv(myInfo,1,MPI_INT,/* from: */ 0,/* labeled: */,0,
    /* not explained here: */&status,MPI_COMM_WORLD);
}
```

Although MPI is sometimes called the 'assembly language of parallel programming', for its perceived difficulty and level of explicitness, it is not all that hard to learn, as evinced by the large number of scientific codes that use it. The main issues that make MPI somewhat intricate to use, are buffer management and blocking semantics.

These issues are related, and stem from the fact that, ideally, data should not be in two places at the same time. Let us briefly consider what happens if processor 1 sends data to processor 2. The safest strategy is for processor 1 to execute the send instruction, and then wait until processor 2 acknowledges that the data was successfully received. This means that processor 1 is temporarily blocked until processor 2 actually executes its receive instruction, and the data has made its way through the network. Alternatively, processor 1 could put its data in a buffer, tell the system to make sure that it gets sent at some point, and later checks to see that the buffer is safe to reuse. This second strategy is called *non-blocking communication*, and it requires the use of a temporary buffer.

2.5.2.3 Collective operations

In the above examples, you saw the `MPI_Allreduce` call, which computed a global sum and left the result on each processor. There is also a local version `MPI_Reduce` which computes the result only on one processor. These calls are examples of *collective operations* or *collectives*. The collectives are:

reduction : each processor has a data item, and these items need to be combined arithmetically with an addition, multiplication, max, or min operation. The result can be left on one processor, or on all, in which case we call this an **allreduce** operation.

broadcast : one processor has a data item that all processors need to receive.

gather : each processor has a data item, and these items need to be collected in an array, without combining them in an operations such as an addition. The result can be left on one processor, or on all, in which case we call this an **allgather**.

scatter : one processor has an array of data items, and each processor receives one element of that array.

all-to-all : each processor has an array of items, to be scattered to all other processors.

We will analyze the cost of collective operations in detail in section 6.3.1.

2.5.2.4 MPI version 1 and 2

The first MPI standard [67] had a number of notable omissions, which are included in the MPI 2 standard [46]. One of these concerned parallel input/output: there was no facility for multiple processes to access the same file, even if the underlying hardware would allow that. A separate project MPI-I/O has now been rolled into the MPI-2 standard.

A second facility missing in MPI, though it was present in PVM [25, 36] which predates MPI, is process management: there is no way to create new processes and have them be part of the parallel run.

Finally, MPI-2 has support for one-sided communication: one process can do a send, without the receiving process actually doing a receive instruction.

2.5.2.5 Non-blocking communication

In a simple computer program, each instruction takes some time to execute, in a way that depends on what goes on in the processor. In parallel programs the situation is more complicated. A send operation, in its simplest form, declares that a certain buffer of data needs to be sent, and program execution will then stop until that buffer has been safely sent and received by another processor. This sort of operation is called a *non-local operation* since it depends on the actions of other processes, and a *blocking communication* operation since execution will halt until a certain event takes place.

Blocking operations have the disadvantage that they can lead to *deadlock*, if two processes wind up waiting for each other. Even without deadlock, they can lead to considerable *idle time* in the processors, as they wait without performing any useful work. On the other hand, they have the advantage that it is clear when the buffer can be reused: after the operation completes, there is a guarantee that the data has been safely received at the other end.

The blocking behaviour can be avoided, at the cost of complicating the buffer semantics, by using *non-blocking operations*. A non-blocking send (`MPI_Isend`) declares that a data buffer needs to be sent, but then does not wait for the completion of the corresponding receive. There is a second operation `MPI_Wait` that will actually block until the receive has been completed. The advantage of this decoupling of sending and blocking is that it now becomes possible to write:

```
ISend(somebuffer,&handle); // start sending, and
    // get a handle to this particular communication
{ ... }  // do useful work on local data
Wait(handle); // block until the communication is completed;
{ ... }  // do useful work on incoming data
```

With a little luck, the local operations take more time than the communication, and you have completely eliminated the communication time.

In addition to non-blocking sends, there are non-blocking receives. A typical piece of code then looks like

```
ISend(sendbuffer,&sendhandle);
IReceive(recvbuffer,&recvhandle);
```

```
{ ... }  // do useful work on local data
Wait(sendhandle); Wait(recvhandle);
{ ... }  // do useful work on incoming data
```

Exercise 2.3. Go back to exercise 1 and give pseudocode that solves the problem using non-blocking sends and receives. What is the disadvantage of this code over a blocking solution?

2.5.3 Parallel languages

One approach to mitigating the difficulty of parallel programming is the design of languages that offer explicit support for parallelism. There are several approaches, and we will see some examples.

- Some languages reflect the fact that many operations in scientific computing are data parallel (section 2.4.1). Languages such as High Performance Fortran (HPF) (section 2.5.3.4) have an *array syntax*, where operations such addition of arrays can be expressed `A = B+C`. This syntax simplifies programming, but more importantly, it specifies operations at an abstract level, so that a lower level can make specific decision about how to handle parallelism. However, the data parallelism expressed in HPF is only of the simplest sort, where the data are contained in regular arrays. Irregular data parallelism is harder; the *Chapel* language (section 2.5.3.6) makes an attempt at addressing this.
- Another concept in parallel languages, not necessarily orthogonal to the previous, is that of Partitioned Global Address Space (PGAS) model: there is only one address space (unlike in the MPI model), but this address space is partitioned, and each partition has affinity with a thread or process. Thus, this model encompasses both SMP and distributed shared memory. The typical PGAS languages, Unified Parallel C (UPC), allows you to write programs that for the most part looks like regular C code. However, by indicating how the major arrays are distributed over processors, the program can be executed in parallel.

2.5.3.1 Discussion

Parallel languages hold the promise of making parallel programming easier, since they make communication operations appear as simple copies or arithmetic operations. However, by doing so they invite the user to write code that may not be efficient, for instance by inducing many small messages.

As an example, consider arrays `a,b` that have been horizontally partitioned over the processors, and that are shifted:

```
for (i=0; i<N; i++)
  for (j=0; j<N/np; j++)
    a[i][j+joffset] = b[i][j+1+joffset]
```

If this code is executed on a shared memory machine, it will be efficient, but a naive translation in the distributed case will have a single number being communicated in each iteration of the `i` loop. Clearly, these can be combined in a single buffer send/receive operation, but compilers are usually unable to make this transformation. As a result, the user is forced to, in effect, re-implement the blocking that needs to be done in an MPI implementation:

```
for (i=0; i<N; i++)
  t[i] = b[i][N/np+joffset]
```

```
  for (i=0; i<N; i++)
    for (j=0; j<N/np-1; j++) {
      a[i][j] = b[i][j+1]
      a[i][N/np] = t[i]
    }
```

On the other hand, certain machines support direct memory copies through global memory hardware. In that case, PGAS languages can be more efficient than explicit message passing, even with physically distributed memory.

2.5.3.2 Unified Parallel C

Unified Parallel C (UPC) [10] is an extension to the C language. Its main source of parallelism is *data parallelism*, where the compiler discovers indepence of operations on arrays, and assigns them to separate processors. The language has an extended array declaration, which allows the user to specify whether the array is partitioned by blocks, or in a *round-robin* fashion.

The following program in UPC performs a vector-vector addition.

```
//vect_add.c
#include <upc_relaxed.h>
#define N 100*THREADS
shared int v1[N], v2[N], v1plusv2[N];
void main() {
  int i;
  for(i=MYTHREAD; i<N; i+=THREADS)
    v1plusv2[i]=v1[i]+v2[i];
}
```

The same program with an explicitly parallel loop construct:

```
//vect_add.c
#include <upc_relaxed.h>
#define N 100*THREADS
shared int v1[N], v2[N], v1plusv2[N];
void main()
{
  int i;
  upc_forall(i=0; i<N; i++; i)
    v1plusv2[i]=v1[i]+v2[i];
}
```

2.5.3.3 Titanium

Titanium is comparable to UPC in spirit, but based on Java rather than on C.

2.5.3.4 High Performance Fortran

High Performance Fortran[3] (HPF) is an extension of Fortran 90 with constructs that support parallel computing, published by the High Performance Fortran Forum (HPFF). The HPFF was convened and chaired by Ken Kennedy of Rice University. The first version of the HPF Report was published in 1993.

Building on the array syntax introduced in Fortran 90, HPF uses a data parallel model of computation to support spreading the work of a single array computation over multiple processors. This allows efficient implementation on both SIMD and MIMD style architectures. HPF features included:

- New Fortran statements, such as FORALL, and the ability to create PURE (side effect free) procedures
- Compiler directives for recommended distributions of array data
- Extrinsic procedure interface for interfacing to non-HPF parallel procedures such as those using message passing
- Additional library routines - including environmental inquiry, parallel prefix/suffix (e.g., 'scan'), data scattering, and sorting operations

Fortran 95 incorporated several HPF capabilities. While some vendors did incorporate HPF into their compilers in the 1990s, some aspects proved difficult to implement and of questionable use. Since then, most vendors and users have moved to OpenMP-based parallel processing. However, HPF continues to have influence. For example the proposed BIT data type for the upcoming Fortran-2008 standard contains a number of new intrinsic functions taken directly from HPF.

2.5.3.5 Co-array Fortran

Co-array Fortran (CAF) is an extension to the Fortran 95/2003 language. The main mechanism to support parallelism is an extension to the array declaration syntax, where an extra dimension indicates the parallel distribution. For instance,

```
real,allocatable,dimension(:,:,:)[:,:] :: A
```

declares an array that is three-dimensional on each processor, and that is distributed over a two-dimensional processor grid.

Communication between processors is now done through copies along the dimensions that describe the processor grid:

```
      COMMON/XCTILB4/ B(N,4)[*]
      SAVE  /XCTILB4/
C
      CALL SYNC_ALL( WAIT=(/IMG_S,IMG_N/) )
      B(:,3) = B(:,1)[IMG_S]
      B(:,4) = B(:,2)[IMG_N]
      CALL SYNC_ALL( WAIT=(/IMG_S,IMG_N/) )
```

The Fortran 2008 standard will include co-arrays.

3. This section quoted from Wikipedia

2.5.3.6 Chapel

Chapel [1] is a new parallel programming language[4] being developed by Cray Inc. as part of the DARPA-led High Productivity Computing Systems program (HPCS). Chapel is designed to improve the productivity of high-end computer users while also serving as a portable parallel programming model that can be used on commodity clusters or desktop multicore systems. Chapel strives to vastly improve the programmability of large-scale parallel computers while matching or beating the performance and portability of current programming models like MPI.

Chapel supports a multithreaded execution model via high-level abstractions for data parallelism, task parallelism, concurrency, and nested parallelism. Chapel's locale type enables users to specify and reason about the placement of data and tasks on a target architecture in order to tune for locality. Chapel supports global-view data aggregates with user-defined implementations, permitting operations on distributed data structures to be expressed in a natural manner. In contrast to many previous higher-level parallel languages, Chapel is designed around a multiresolution philosophy, permitting users to initially write very abstract code and then incrementally add more detail until they are as close to the machine as their needs require. Chapel supports code reuse and rapid prototyping via object-oriented design, type inference, and features for generic programming.

Chapel was designed from first principles rather than by extending an existing language. It is an imperative block-structured language, designed to be easy to learn for users of C, C++, Fortran, Java, Perl, Matlab, and other popular languages. While Chapel builds on concepts and syntax from many previous languages, its parallel features are most directly influenced by ZPL, High-Performance Fortran (HPF), and the Cray MTA's extensions to C and Fortran.

Here is vector-vector addition in Chapel:

```
const BlockDist= newBlock1D(bbox=[1..m], tasksPerLocale=...);
const ProblemSpace: domain(1, 64)) distributed BlockDist = [1..m];
varA, B, C: [ProblemSpace] real;
forall(a, b, c) in(A, B, C) do
  a = b + alpha * c;
```

2.5.3.7 Fortress

Fortress [6] is a programming language developed by Sun Microsystems. Fortress[5] aims to make parallelism more tractable in several ways. First, parallelism is the default. This is intended to push tool design, library design, and programmer skills in the direction of parallelism. Second, the language is designed to be more friendly to parallelism. Side-effects are discouraged because side-effects require synchronization to avoid bugs. Fortress provides transactions, so that programmers are not faced with the task of determining lock orders, or tuning their locking code so that there is enough for correctness, but not so much that performance is impeded. The Fortress looping constructions, together with the library, turns "iteration" inside out; instead of the loop specifying how the data is accessed, the data structures specify how the loop is run, and aggregate data structures are designed to break into large parts that can be effectively scheduled for parallel execution. Fortress also includes features from other languages intended to generally help productivity – test code and methods, tied to the code under test; contracts

4. This section quoted from the Chapel homepage.
5. This section quoted from the Fortress homepage.

that can optionally be checked when the code is run; and properties, that might be too expensive to run, but can be fed to a theorem prover or model checker. In addition, Fortress includes safe-language features like checked array bounds, type checking, and garbage collection that have been proven-useful in Java. Fortress syntax is designed to resemble mathematical syntax as much as possible, so that anyone solving a problem with math in its specification can write a program that can be more easily related to its original specification.

2.5.3.8 X10

X10 is an experimental new language currently under development at IBM in collaboration with academic partners. The X10 effort is part of the IBM PERCS project (Productive Easy-to-use Reliable Computer Systems) in the DARPA program on High Productivity Computer Systems. The PERCS project is focused on a hardware-software co-design methodology to integrate advances in chip technology, architecture, operating systems, compilers, programming language and programming tools to deliver new adaptable, scalable systems that will provide an order-of-magnitude improvement in development productivity for parallel applications by 2010.

X10 aims to contribute to this productivity improvement by developing a new programming model, combined with a new set of tools integrated into Eclipse and new implementation techniques for delivering optimized scalable parallelism in a managed runtime environment. X10 is a type-safe, modern, parallel, distributed object-oriented language intended to be very easily accessible to Java(TM) programmers. It is targeted to future low-end and high-end systems with nodes that are built out of multi-core SMP chips with non-uniform memory hierarchies, and interconnected in scalable cluster configurations. A member of the Partitioned Global Address Space (PGAS) family of languages, X10 highlights the explicit reification of locality in the form of places; lightweight activities embodied in async, future, foreach, and ateach constructs; constructs for termination detection (finish) and phased computation (clocks); the use of lock-free synchronization (atomic blocks); and the manipulation of global arrays and data structures.

2.5.3.9 Linda

As should be clear by now, the treatment of data is by far the most important aspect of parallel programing, far more important than algorithmic considerations. The programming system *Linda* [37], also called a *coordination language*, is designed to address the data handling explicitly. Linda is not a language as such, but can, and has been, incorporated into other languages.

The basic concept of Linda is *tuple space*: data is added to a pool of globally accessible information by adding a label to it. Processes then retrieve data by their label, and without needing to know which processes added them to the tuple space.

Linda is aimed primarily at a different computation model than is relevant for High-Performance Computing (HPC): it addresses the needs of asynchronous communicating processes. However, is has been used for scientific computation [22]. For instance, in parallel simulations of the heat equation (section 4.3), processors can write their data into tuple space, and neighbouring processes can retrieve their *ghost region* without having to know its provenance. Thus, Linda becomes one way of implementing *one-sided communication*.

2.5.4 OS-based approaches

It is possible to design an architecture with a shared address space, and let the data movement be handled by the operating system. The Kendall Square computer [4] had an architecture name 'all-cache', where no data was natively associated with any processor. Instead, all data was considered to be cached on a processor, and moved through the network on demand, much like data is moved from main memory to cache in a regular CPU. This idea is analogous to the numa support in current SGI architectures.

2.5.5 Latency hiding

Communication between processors is typically slow, slower than data transfer from memory on a single processor, and much slower than operating on data. For this reason, it is good to think about the relative volumes of network traffic versus 'useful' operations when designing a parallel program. Another way of coping with the relative slowness of communication is to arrange the program so that the communication actually happens while some computation is going on. This is referred to as *overlapping computation with communication*.

For example, consider the parallel execution of a matrix-vector product $y = Ax$ (there will be further discussion of this operation in section 6.2). Assume that the vectors are distributed, so each processor p executes

$$\forall_{i\in I_p}\colon y_i = \sum_j a_{ij}x_j.$$

Since x is also distributed, we can write this as

$$\forall_{i\in I_p}\colon y_i = \left(\sum_{j \text{ local}} + \sum_{j \text{ not local}} \right) a_{ij}x_j.$$

This scheme is illustrated in figure 2.4. We can now proceed as follows:

- Start the transfer of non-local elements of x;
- Operate on the local elements of x while data transfer is going on;
- Make sure that the transfers are finished;
- Operate on the non-local elements of x.

Of course, this scenario presupposes that there is software and hardware support for this overlap. MPI allows for this, through so-called *asynchronous communication* or *non-blocking comumnication* routines. This does not immediately imply that overlap will actually happen, since hardware support is an entirely separate question.

2.6 Topologies

If a number of processors are working together on a single task, most likely they need to communicate data. For this reason there needs to be a way for data to make it from any processor to any other. In this section we will discuss some of the possible schemes to connect the processors in a parallel machine.

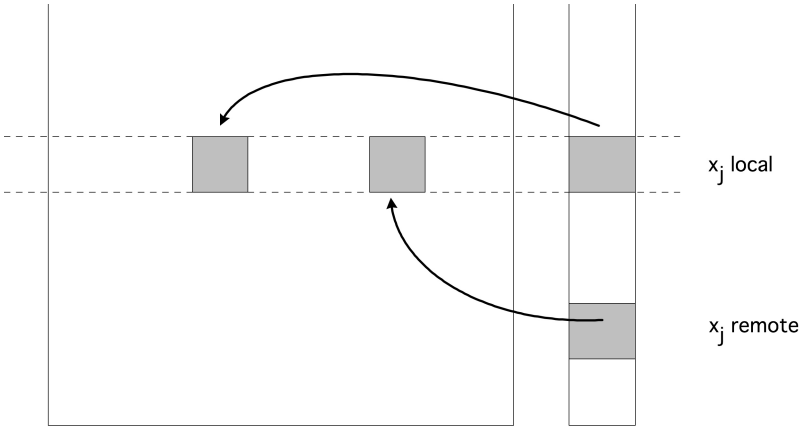


Figure 2.4: The parallel matrix-vector product with a blockrow distribution.

In order to get an appreciation for the fact that there is a genuine problem here, consider two simple schemes that do not 'scale up':

- Ethernet is a connection scheme where all machines on a network are on a single cable[6]. If one machine puts a signal on the wire to send a message, and another also wants to send a message, the latter will detect that the sole available communication channel is occupied, and it will wait some time before retrying its send operation. Receiving data on ethernet is simple: messages contain the address of the intended recipient, so a processor only has to check whether the signal on the wire is intended for it.
  The problems with this scheme should be clear. The capacity of the communication channel is finite, so as more processors are connected to it, the capacity available to each will go down. Because of the scheme for resolving conflicts, the average delay before a message can be started will also increase[7].
- In a *fully connected* configuration, each processor has one wire for the communications with each other processor. This scheme is perfect in the sense that messages can be sent in the minimum amount of time, and two messages will never interfere with each other. The amount of data that can be sent from one processor is no longer a decreasing function of the number of processors; it is in fact an increasing function, and if the network controller can handle it, a processor can even engage in multiple simultaneous communications.
  The problem with this scheme is of course that the design of the network interface of a processor is no longer fixed: as more processors are added to the parallel machine, the network interface gets more connecting wires. The network controller similarly becomes more complicated, and the cost of the machine increases faster than linearly in the number of processors.

In this section we will see a number of schemes that *can* be increased to large numbers of processors.

6. We are here describing the original design of Ethernet. With the use of switches, especially in an HPC context, this description does not really apply anymore.
7. It was initially thought that ethernet would be inferior to other solutions such as IBM's 'token ring'. It takes fairly sophisticated statistical analysis to prove that it works a lot better than was naively expected.

2.6.1 Some graph theory

The network that connects the processors in a parallel computer can conveniently be described with some elementary graph theory concepts. We describe the parallel machine with a graph where each processor is a node, and two nodes are connected[8] if there is a direct connection between them.

We can then analyze two important concepts of this graph.

First of all, the *degree* of a node in a graph is the number of other nodes it is connected to. With the nodes representing processors, and the edges the wires, it is clear that a high degree is not just desirable for efficiency of computing, but also costly from an engineering point of view. We assume that all processors have the same degree.

Secondly, a message traveling from one processor to another, through one or more intermediate nodes, will most likely incur some delay at each intermediate node. For this reason, the *diameter* of the graph is important. The diameter is defined as the maximum shortest distance, counting numbers of wires, between any two processors:

$$d(G) = \max_{i,j} |\text{shortest path between } i \text{ and } j|.$$

If d is the diameter, and if sending a message over one wire takes unit time, this means a message will always arrive in at most time d.

Exercise 2.4. Find a relation between the number of processors, their degree, and the diameter of the connectivity graph.

In addition to the question 'how long will a message from processor A to processor B take', we often worry about conflicts between two simultaneous messages: is there a possibility that two messages, under way at the same time, will need to use the same network link? This sort of conflict is called *congestion* or *contention*. Clearly, the more links a parallel comupter has, the smaller the chance of congestion.

A precise way to describe the likelihood of congestion, is to look at the *bisection width*. This is defined as the minimum number of links that have to be removed to partition the processor graph into two unconnected graphs. For instance, consider processors connected as a linear array, that is, processor P_i is connected to P_{i-1} and P_{i+1}. In this case the bisection width is 1.

The bisection width w describes how many messages can, guaranteed, be under way simultaneously in a parallel computer. Proof: take w sending and w receiving processors. The w paths thus defined are disjoint: if they were not, we could separate the processors into two groups by removing only $w - 1$ links.

In practice, of course, more than w messages can be under way simultaneously. For instance, in a linear array, which has $w = 1$, $P/2$ messages can be sent and received simultaneously if all communication is between neighbours, and if a processor can only send or receive, but not both, at any one time. If processors can both send and receive simultaneously, P messages can be under way in the network.

Bisection width also describes *redundancy* in a network: if one or more connections are malfunctioning, can a message still find its way from sender to receiver?

8. We assume that connections are symmetric, so that the network is an *undirected graph*.

While bisection width is a measure express as a number of wires, in practice we care about the capacity through those wires. The relevant concept here is *bisection bandwidth*: the bandwidth across the bisection width, which is the product of the bisection width, and the capacity (in bits per second) of the wires. Bisection bandwidth can be considered as a measure for the bandwidth that can be attained if an arbitrary half of the processors communicates with the other half. Bisection bandwidth is a more realistic measure than the *aggregate bandwidth* which is sometimes quoted and which is defined as the total data rate if every processor is sending: the number of processors times the bandwidth of a connection times the number of simultaneous sends a processor can perform. This can be quite a high number, and it is typically not representative of the communication rate that is achieved in actual applications.

2.6.2 Linear arrays and rings

A simple way to hook up multiple processors is to connect them in a *linear array*: every processor has a number i, and processor P_i is connected to P_{i-1} and P_{i+1}. The first and last processor are possible exceptions: if they are connected to each other, we call the architecture a *ring network*.

This solution requires each processor to have two network connections, so the design is fairly simple.

Exercise 2.5. What is the bisection width of a linear array? Of a ring?

Exercise 2.6. With the limited connections of a linear array, you may have to be clever about how to program parallel algorithms. For instance, consider a 'broadcast' operation: processor 0 has a data item that needs to be sent to every other processor.

We make the following simplifying assumptions:

- a processor can send any number of messages simultaneously,
- but a wire can can carry only one message at a time; however,
- communication between any two processors takes unit time, regardless the number of processors in between them.

In a fully connected network you can simply write

for $i = 1 \ldots N-1$:

  send the message to processor i

in a fully connected network this means that the operation is done in one step.

Now consider a linear array. Show that, even with this unlimited capacity for sending, the above algorithm runs into trouble because of congestion.

Find a better way to organize the send operations. Hint: pretend that your processors are connected as a binary tree. Assume that there are $N = 2^n$ processors. Show that the broadcast can be done in $\log N$ stages, and that processors only need to be able to send a single message simultaneously.

This exercise is an example of *embedding* a 'logical' communication pattern in a physical one.

2.6.3 2D and 3D arrays

A popular design for parallel computers is to organize the processors in a two-dimensional or three-dimensional *cartesian mesh*. This means that every processor has a coordinate (i, j) or (i, j, k), and it is connected to its neigh-

bours in all coordinate directions. The processor design is still fairly simple: the number of network connections (the degree of the connectivity graph) is twice the number of space dimensions (2 or 3) of the network.

It is a fairly natural idea to have 2D or 3D networks, since the world around us is three-dimensional, and computers are often used to model real-life phenomena. If we accept for now that the physical model requires *nearest neighbour* type communications (which we will see is the case in section 4.2.2.2), then a mesh computer is a natural candidate for running physics simulations.

Exercise 2.7. What is the diameter of a 3D cube of processors? What is the bisection width? How does that change if you add wraparound torus connections?

Exercise 2.8. Your parallel computer has its processors organized in a 2D grid. The chip manufacturer comes out with a new chip with same clock speed that is dual core instead of single core, and that will fit in the existing sockets. Critique the following argument: "the amount work per second that can be done (that does not involve communication) doubles; since the network stays the same, the bisection bandwidth also stays the same, so I can reasonably expect my new machine to become twice as fast[9]."

2.6.4 Hypercubes

Above we gave a hand-waving argument for the suitability of mesh-organized processors, based on the prevalence of nearest neighbour communications. However, sometimes sends and receives between arbitrary processors occur. One example of this is the above-mentioned broadcast. For this reason, it is desirable to have a network with a smaller diameter than a mesh. On the other hand we want to avoid the complicated design of a fully connected network.

A good intermediate solution is the *hypercube* design. An n-dimensional hypercube computer has 2^n processors, with each processor connected to one other in each dimension; see figure 2.5.

An easy way to describe this is to give each processor an address consisting of d bits: we give each node of a hypercube a number that is the bit pattern describing its location in the cube; see figure 2.6.

With this numbering scheme, a processor is then connected to all others that have an address that differs by exactly one bit. This means that, unlike in a grid, a processor's neighbours do not have numbers that differ by 1 or $\sqrt{P}$, but by $1, 2, 4, 8, \ldots$.

The big advantages of a hypercube design are the small diameter and large capacity for traffic through the network.

Exercise 2.9. What is the diameter of a hypercube? What is the bisection width?

One disadvantage is the fact that the processor design is dependent on the total machine size. In practice, processors will be designed with a maximum number of possible connections, and someone buying a smaller machine then will be paying for unused capacity. Another disadvantage is the fact that extending a given machine can only be done by doubling it: other sizes than 2^p are not possible.

9. With the numbers one and two replaced by higher numbers, this is actually not a bad description of current trends in processor design.

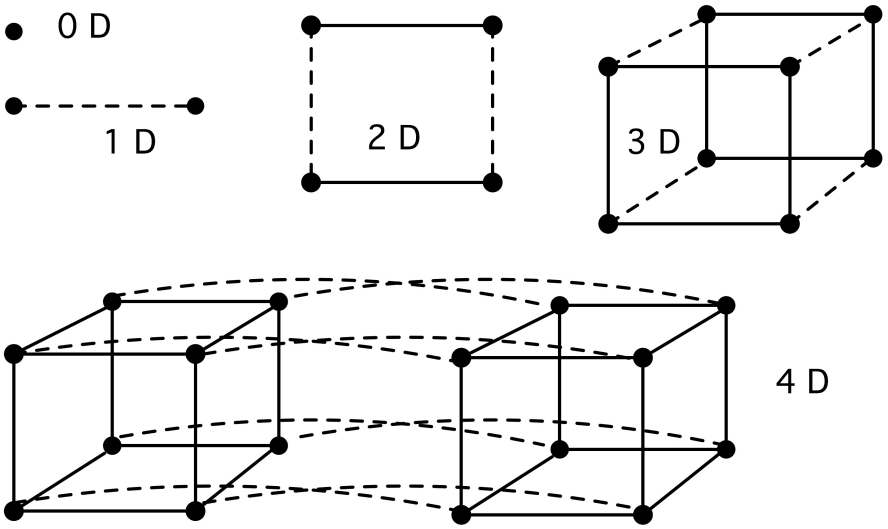


Figure 2.5: Hypercubes

Exercise 2.10. Consider the parallel summing example of section 2.1, and give the execution time of a parallel implementation, first on a linear array, then on a hypercube. Assume that sending one number from a processor to a neighbour takes time α and that a floating point operation takes time γ. Show that on a linear array the algorithm gives at best a factor speedup over the sequential algorithm; show that the theoretical speedup from the example is attained (up to a factor) for the implementation on a hypercube.

2.6.4.1 Embedding grids in a hypercube

Above we made the argument that mesh-connected processors are a logical choice for many applications that model physical phenomena. How is that for hypercubes? The answer is that a hyercube has enough connections that it can simply pretend to be a mesh by ignoring certain connections. However, we can not use the obvious numbering of nodes as in figure 2.6. For instance, node 1 is directly connected to node 0, but has a distance of 2 to node 2. The right neighbour of node 3 in a ring, node 4, even has the maximum distance of 3 in this hypercube. To explain how we can embed a mesh in a hypercube, we first show that it's possible to walk through a hypercube, touching every corner exactly once.

The basic concept here is a (binary reflected) *Gray code* [44]. This is a way of ordering the binary numbers $0 \ldots 2^d - 1$ as $g_0, \ldots g_{2^d-1}$ so that g_i and g_{i+1} differ in only one bit. Clearly, the ordinary binary numbers do not satisfy this: the binary representations for 1 and 2 already differ in two bits. Why do Gray codes help us? Well, since g_i and g_{i+1} differ only in one bit, it means they are the numbers of nodes in the hypercube that are directly connected.

Figure 2.7 illustrates how to construct a Gray code. The procedure is recursive, and can be described informally as

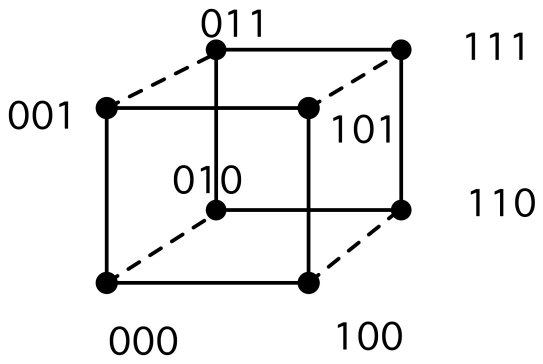


Figure 2.6: Numbering of the nodes of a hypercube

'divide the cube into two subcubes, number the one subcube, cross over to the other subcube, and number its nodes in the reverse order of the first one'.

1D Gray code:

$$0 \quad 1$$

2D Gray code:

$$\begin{array}{rccccc} \text{1D code and reflection:} & 0 & 1 & \vdots & 1 & 0 \\ \text{append 0 and 1 bit:} & 0 & 0 & \vdots & 1 & 1 \end{array}$$

3D Gray code:

$$\begin{array}{rcccccccc} \text{2D code and reflection:} & 0 & 1 & 1 & 0 & \vdots & 0 & 1 & 1 & 0 \\ & 0 & 0 & 1 & 1 & \vdots & 1 & 1 & 0 & 0 \\ \text{append 0 and 1 bit:} & 0 & 0 & 0 & 0 & \vdots & 1 & 1 & 1 & 1 \end{array}$$

Figure 2.7: Gray codes

Since a Gray code offers us a way to embed a one-dimensional 'mesh' into a hypercube, we can now work our way up.

Exercise 2.11. Show how a square mesh of 2^{2d} nodes can be embedded in a hypercube by appending the bit patterns of the embeddings of two 2^d node cubes. How would you accomodate a mesh of $2^{d_1+d_2}$ nodes? A three-dimensional mesh of $2^{d_1+d_2+d_3}$ nodes?

2.6.5 Switched networks

Above, we briefly discussed fully connected processors. They are impractical if the connection is made by making a large number of wires between all the processors. There is another possibility, however, by connecting all the processors to a *switch* or switching network. Some popular network designs are the *crossbar*, the *butterfly exchange*, and the *fat tree*.

Switching networks are made out of switching elements, each of which have a small number (up to about a dozen)

of inbound and outbound links. By hooking all processors up to some switching element, and having multiple stages of switching, it then becomes possible to connect any two processors by a path through the network.

2.6.5.1 Cross bar

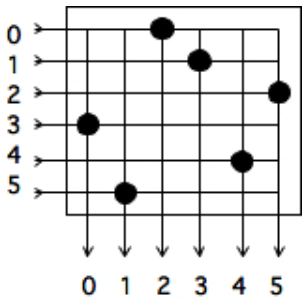


Figure 2.8: A simple cross bar connecting 6 inputs to 6 outputs

The simplest switching network is a cross bar, an arrangement of n horizontal and vertical lines, with a switch element on each intersection that determines whether the lines are connected; see figure 2.8. If we designate the horizontal lines as inputs the vertical as outputs, this is clearly a way of having n inputs be mapped to n outputs. Every combination of inputs and outputs (sometimes called a 'permutation') is allowed.

2.6.5.2 Butterfly exchange

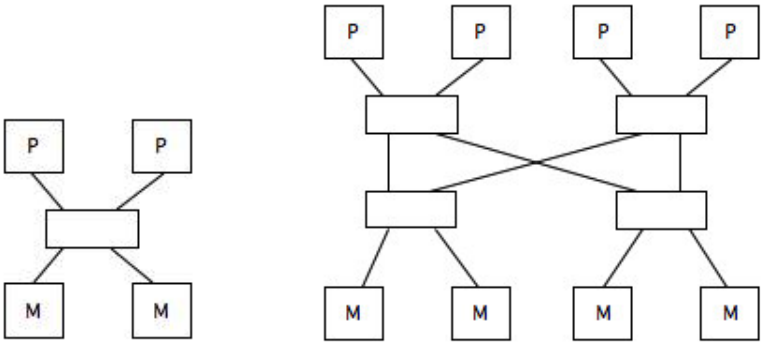


Figure 2.9: A butterfly exchange network for two and four processors/memories

Butterfly exchanges are typically built out of small switching elements, and they have multiple stages; as the number of processors grows, the number of stages grows with it. As you can see in figure 2.10, butterfly exchanges allow several processors to access memory simultaneously. Also, their access times are identical, so exchange networks are a way of implementing a UMA architecture; see section 2.3.1.

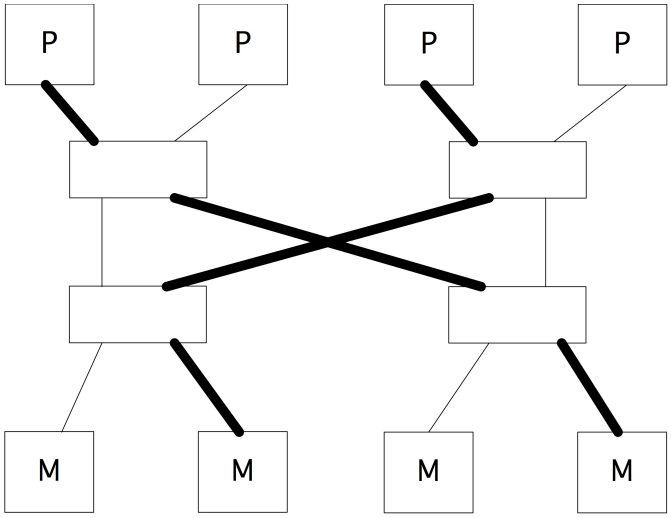


Figure 2.10: Two independent routes through a butterfly exchange network

Exercise 2.12. For both the simple cross bar and the butterfly exchange, the network needs to be expanded as the number of processors grows. Give the number of wires (of some unit length) and the number of switching elements that is needed in both cases to connect n processors and memories. What is the time that a data packet needs to go from memory to processor, expressed in the unit time that it takes to traverse a unit length of wire and the time to traverse a switching element?

2.6.5.3 Fat-trees

If we were to connect switching nodes like a tree, there would be a big problem with congestion close to the root since there are only two wires attached to the root note. Say we have a k-level tree, so there are 2^k leaf nodes. If all leaf nodes in the left subtree try to communicate with nodes in the right subtree, we have 2^{k-1} messages going through just one wire into the root, and similarly out through one wire. A fat-tree is a tree network where each level has the same total bandwidth, so that this congestion problem does not occur: the root will actually have 2^{k-1} incoming and outgoing wires attached [45]; see figure 2.11. The first successful computer architecture based on a fat-tree was the Connection Machines CM5.

In fat-trees, as in other switching networks, each message carries its own routing information. Since in a fat-tree the choices are limited to going up a level, or switching to the other subtree at the current level, a message needs to carry only as many bits routing information as there are levels, which is $\log_2 n$ for n processors.

The theoretical exposition of fat-trees in [62] shows that fat-trees are optimal in some sense: it can deliver messages as fast (up to logarithmic factors) as any other network that takes the same amount of space to build. The underlying assumption of this statement is that switches closer to the root have to connect more wires, therefore take more components, and correspondingly are larger.

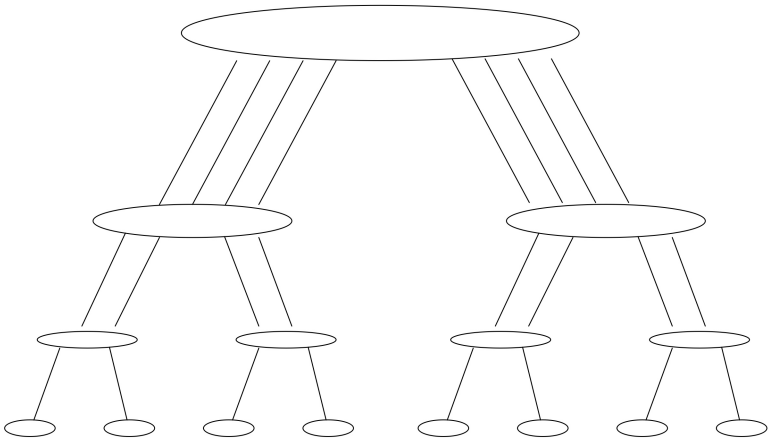


Figure 2.11: A fat tree with a three-level interconnect

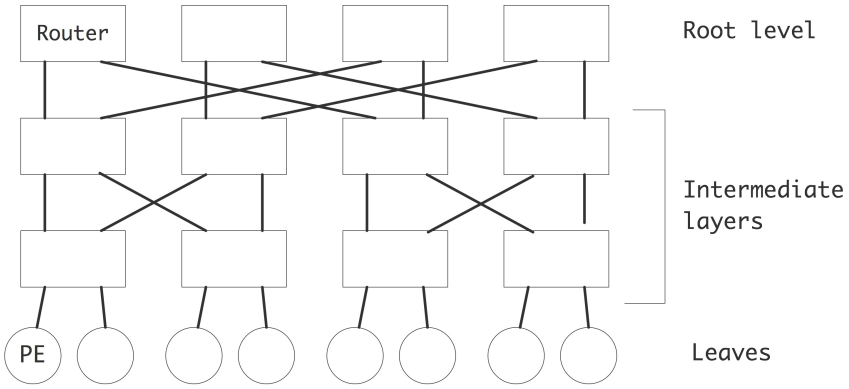


Figure 2.12: A fat-tree built from simple switching elements

This argument, while theoretically interesting, is of no practical significance, as the physical size of the network hardly plays a role in the biggest currently available computers that use fat-tree interconnect. For instance, in the Ranger supercomputer of The University of Texas at Austin, the fat-tree switch connects 60,000 cores, yet takes less than 10 percent of the floor space.

A fat tree, as sketched above, would be costly to build, since for every next level a new, bigger, switch would have to be designed. In practice, therefore, a network with the characteristics of a fat-tree is constructed from simple switching elements; see figure 2.12. This network is equivalent in its bandwidth and routing possibilities to a fat-tree. Routing algorithms will be slightly more complicated: in a fat-tree, a data packet can go up in only one way, but here a packet has to know to which of the two higher switches to route.

This type of switching network is one case of a *Clos network* [21].

2.6.6 Bandwidth and latency

The statement above that sending a message can be considered a unit time operation, is of course unrealistic. A large message will take longer to transmit than a short one. There are two concepts to arrive at a more realistic description of the transmission process; we have already seen this in section 1.2.2 in the context of transferring data between cache levels of a processor.

latency Setting up a communication between two processors takes an amount of time that is independent of the message size. The time that this takes is known as the *latency* of a message. There are various causes for this delay.

- The two processors engage in 'hand-shaking', to make sure that the recipient is ready, and that appropriate buffer space is available for receiving the message.
- The message needs to be encoded for transmission by the sender, and decoded by the receiver.
- The actual transmission may take time: parallel computers are often big enough that, even at light-speed, the first byte of a message can take hundreds of cycles to traverse the distance between two processors.

bandwidth After a transmission between two processors has been initiated, the main number of interest is the number of bytes per second that can go through the channel. This is known as the *bandwidth*. The bandwidth can usually be determined by the *channel rate*, the rate at which a physical link can deliver bits, and the *channel width*, the number of physical wires in a link. The channel width is typically a multiple of 16, usually 64 or 128. This is also expressed by saying that a channel can send one or two 8-byte words simultaneously.

Bandwidth and latency are formalized in the expression

$$T(n) = \alpha + \beta n$$

for the transmission time of an n-byte message. Here, α is the latency and β is the time per byte, that is, the inverse of bandwidth.

2.7 Theory

There are two important reasons for using a parallel computer: to have access to more memory or to obtain higher performance. It is easy to characterize the gain in memory, as the total memory is the sum of the individual memories. The speed of a parallel computer is harder to characterize. A simple approach is to let the same program run on a single processor, and on a parallel machine with p processors, and to compare runtimes.

With T_1 the execution time on a single processor and T_p the time on p processors, we define the *speedup* as $S_p = T_1/T_p$. (Sometimes T_1 is defined as 'the best time to solve the problem on a single processor', which allows for using a different algorithm on a single processor than in parallel.) In the ideal case, $T_p = T_1/p$, but in practice we don't expect to attain that, so $S_P \leq p$. To measure how far we are from the ideal speedup, we introduce the *efficiency* $E_p = S_p/p$. Clearly, $0 < E_p \leq 1$.

There is a practical problem with this definition: a problem that can be solved on a parallel machine may be too large to fit on any single processor. Conversely, distributing a single processor problem over many processors may give a distorted picture since very little data will wind up on each processor. Below we will discuss more realistic measures of speed-up.

There are various reasons why the actual speed is less than P. For one, using more than one processors necessitates communication, which is overhead that was not part of the original computation. Secondly, if the processors do not have exactly the same amount of work to do, they may be idle part of the time (this is known as *load unbalance*), again lowering the actually attained speedup. Finally, code may have sections that are inherently sequential.

Communication between processors is an important source of a loss of efficiency. Clearly, a problem that can be solved without communication will be very efficient. Such problems, in effect consisting of a number of completely independent calculations, is called *embarassingly parallel*; it will have close to a perfect speedup and efficiency.

Exercise 2.13. The case of speedup larger than the number of processors is called *superlinear speedup*. Give a theoretical argument why this can never happen.

In practice, superlinear speedup can happen. For instance, suppose a problem is too large to fit in memory, and a single processor can only solve it by swapping data to disc. If the same problem fits in the memory of two processors, the speedup may well be larger than 2 since disc swapping no longer occurs. Having less, or more localized, data may also improve the cache behaviour of a code.

2.7.1 Amdahl's law

One reason for less than perfect speedup is that parts of a code can be inherently sequential. This limits the parallel efficiency as follows. Suppose that 5% of a code is sequential, then the time for that part can not be reduced, no matter how many processors are available. Thus, the speedup on that code is limited to a factor of 20. This phenomenon is known as *Amdahl's Law* [11], which we will now formulate.

Let F_p be the *parallel fraction* and F_s be the *parallel fraction* (or more strictly: the 'parallelizable' fraction) of a code, respectively. Then $F_p + F_s = 1$. The parallel execution time T_p is the sum of the part that is sequential $T_1 F_s$ and the part that can be parallelized $T_1 F_p/P$:

$$T_P = T_1(F_s + F_p/P). \tag{2.2}$$

As the number of processors grows $P \to \infty$, the parallel execution time now approaches that of the sequential fraction of the code: $T_P \downarrow T_1 F_s$. We conclude that speedup is limited by $S_P \leq 1/F_s$ and efficiency is a decreasing function $E \sim 1/P$.

The sequential fraction of a code can consist of things such as I/O operations. However, there are also parts of a code that in effect act as sequential. Consider a program that executes a single loop, where all iterations can be computed independently. Clearly, this code is easily parallelized. However, by splitting the loop in a number of parts, one per processor, each processor now has to deal with loop overhead: calculation of bounds, and the test for completion. This overhead is replicated as many times as there are processors. In effect, loop overhead acts as a sequential part of the code.

Exercise 2.14. Investigate the implications of Amdahls's law: if the number of processors P increases, how does the parallel fraction of a code have to increase to maintain a fixed efficiency?

2.7.1.1 Amdahl's law with communication overhead

In a way, Amdahl's law, sobering as it is, is even optimistic. Parallelizing a code will give a certain speedup, but it also introduces *communication overhead* that will lower the speedup attained. Let us refine our model of equation (2.2) (see [60, p. 367]):

$$T_p = T_1(F_s + F_p/P) + T_c,$$

where T_c is a fixed communication time.

To assess the influence of this communication overhead, we assume that the code is fully parallelizable, that is, $f_p = 1$. We then find that

$$S_p = \frac{T_1}{T_1/p + T_c}.$$

For this to be close to p, we need $T_c \ll T_1/p$ or $p \ll T_1/T_c$. In other words, the number of processors should not grow beyond the ratio of scalar execution time and communication overhead.

2.7.1.2 Gustafson's law

Amdahl's law was thought to show that large numbers of processors would never pay off. However, the implicit assumption in Amdahl's law is that there is a fixed computation which gets executed on more and more processors. In practice this is not the case: typically there is a way of scaling up a problem, and one tailors the size of the problem to the number of available processors.

A more realistic assumption would be to say that the sequential fraction is independent of the problem size, and the parallel fraction can be arbitrarily extended. To formalize this, instead of starting with the execution time of the sequential program, let us start with the execution time of the parallel program, and say that

$$T_p = f_s + f_p \equiv 1,$$

and the execution time of this problem on a sequential processor would then be

$$T_1 = f_s + p \cdot f_p.$$

This gives us a speedup of

$$S_p = \frac{T_1}{T_p} = \frac{f_s + p \cdot f_p}{f_s + f_p} = f_s + p \cdot f_p = p - (p-1) \cdot f_s.$$

That is, speedup is now a function that decreases from p, linearly with p.

2.7.2 Scalability

Above, we remarked that splitting a given problem over more and more processors does not make sense: at a certain point there is just not enough work for each processor to operate efficiently. Instead, in practice, users of a parallel code will either choose the number of processors to match the problem size, or they will solve a series of increasingly larger problems on correspondingly growing numbers of processors. In both cases it is hard to talk about speedup. Instead, the concept of *scalability* is used.

We distinguish two types of scalability. So-called *strong scalability* is in effect the same as speedup, discussed above. We say that a program shows strong scalability if, partitioned over more and more processors, it shows perfect or near perfect speedup. Typically, one encounters statements like 'this problem scales up to 500 processors', meaning that up to 500 processors the speedup will not noticeably decrease from optimal.

More interesting, *weak scalability* is a more vaguely defined term. It describes that, as problem size and number of processors grow in such a way that the amount of data per processor stays constant, the speed in operations per second of each processor also stays constant. This measure is somewhat hard to report, since the relation between the number of operations and the amount of data can be complicated. If this relation is linear, one could state that the amount of data per processor is kept constant, and report that parallel execution time is constant as the number of processors grows.

Scalability depends on the way an algorithm is parallelized, in particular on the way data is distributed. In section 6.3 you will find an analysis of the matrix-vector product operation: distributing a matrix by block rows turns out not to be scalable, but a two-dimensional distribution by submatrices is.

2.8 Load balancing

In much of this chapter, we assumed that a problem could be perfectly divided over processors, that is, a processor would always be performing useful work, and only be *idle* because of latency in communication. In practice, however, a processor may be idle because it is waiting for a message, and the sending processor has not even reached the send instruction in its code. Such a situation, where one processor is working and another is idle, is described as *load unbalance*: there is no intrinsic reason for the one processor to be idle, and it could have been working if we had distributed the work load differently.

2.8.1 Load balancing of independent tasks

Let us consider the case of a job that can be partitioned into independent tasks, for instance computing the pixels of a *Mandelbrot set* picture, where each pixel is set according to a mathematical function that does not depend on surrounding pixels. If we could predict the time it would take to draw an arbitrary part of the picture, we could make a perfect division of the work and assign it to the processors. This is known as *static load balancing*.

More realistically, we can not predict the running time of a part of the job perfectly, and we use an *over-decomposition* of the work: we divide the work in more tasks than there are processors. These tasks are then assigned to a *work pool*, and processors take the next job from the pool whenever they finish a job. This is known as *dynamic load balancing*. Many graph and combinatorial problems can be approached this way.

2.8.2 Load balancing as graph problem

A parallel computation can be formulated as a graph (see Appendix A.6 for an introduction to graph theory) where the processors are the vertices, and there is an edge between two vertices if their processors need to communicate at some point. Such a graph is often derived from an underlying graph of the problem being solved. Let us consider for example the matrix-vector product $y = Ax$ where A is a sparse matrix, and look in detail at the processor that is computing y_i for some i. The statement $y_i \leftarrow y_i + A_{ij}x_j$ implies that this processor will need the value x_j, so, if this variable is on a different processor, it needs to be sent over.

We can formalize this: Let the vectors x and y be distributed disjointly over the processors, and define uniquely $P(i)$ as the processor that owns index i. Then there is an edge (P, Q) if there is a nonzero element a_{ij} with $P = P(i)$ and $Q = P(j)$. This graph is undirected if the matrix is *structurally symmetric*, that is $a_{ij} \neq 0 \Leftrightarrow a_{ji} \neq 0$.

The distribution of indices over the processors now gives us vertex and edge weights: a processor has a vertex weight that is the number of indices owned by it; an edge (P, Q) has a weight that is the number of vector components that need to be sent from Q to P, as described above.

The *load balancing* problem can now be formulated as follows:

> Find a partitioning $\mathbb{P} = \cup_i \mathbb{P}_i$, such the variation in vertex weights is minimal, and simultaneously the edge weights are as low as possible.

Minimizing the variety in vertex weights implies that all processor have approximately the same amount of work. Keeping the edge weights low means that the amount of communication is low. These two objectives need not be satisfiable at the same time: some trade-off is likely.

Exercise 2.15. Consider the limit case where processors are infinitely fast and bandwidth between processors is also unlimited. What is the sole remaining factor determining the runtime? What graph problem do you need to solve now to find the optimal load balance? What property of a sparse matrix gives the worst case behaviour?

2.9 The TOP500 List

There are several informal ways of measuring just 'how big' a computer is. The most popular is the TOP500 list, maintained at `top500.org`, which records a computer's performance on the *LINPACK benchmark*. LINPACK is a package for linear algebra operations, and no longer in use, since it has been superseded by *Lapack*. The benchmark operation is the solution of a (square, nonsingular, dense) linear system through LU factorization with partial pivoting, with subsequent forward and backward solution.

The LU factorization operation is one that has great opportunity for cache reuse, since it is based on the matrix-matrix multiplication kernel discussed in section 1.4.1. As a result, the LINPACK benchmark is likely to run at a substantial fraction of the peak speed of the machine. Another way of phrasing this is to say that the LINPACK benchmark is CPU-bound.

Typical efficiency figures are between 60 and 90 percent. However, it should be noted that many scientific codes do not feature the dense linear solution kernel, so the performance on this benchmark is not indicative of the

performance on a typical code. Linear system solution through iterative methods (section 5.5) is much less efficient, being dominated by the bandwidth between CPU and memory ('bandwidth bound').

One implementation of the LINPACK benchmark that is often used is 'High-Performance LINPACK' (`http://www.netlib.org/benchmark/hpl/`), which has several parameters such as blocksize that can be chosen to tune the performance.

Chapter 3

Computer Arithmetic

Of the various types of data that one normally encounters, the ones we are concerned with in the context of scientific computing are the numerical types: integers (or whole numbers) $\ldots, -2, -1, 0, 1, 2, \ldots$, real numbers $0, 1, -1.5, 2/3, \sqrt{2}, \log 10, \ldots$, and complex numbers $1 + 2i, \sqrt{3} - \sqrt{5}i, \ldots$. Computer memory is organized to give only a certain amount of space to represent each number, in multiples of *bytes*, each containing 8 *bits*. Typical values are 4 bytes for an integer, 4 or 8 bytes for a real number, and 8 or 16 bytes for a complex number.

Since only a certain amount of memory is available to store a number, it is clear that not all numbers of a certain type can be stored. For instance, for integers only a range is stored. In the case of real numbers, even storing a range is not possible since any interval $[a, b]$ contains infinitely many numbers. Therefore, any *representation of real numbers* will cause gaps between the numbers that are stored. As a result, any computation that results in a number that is not representable will have to be dealt with by issuing an error or by approximating the result. In this chapter we will look at the ramifications of such approximations of the 'true' outcome of numerical calculations.

3.1 Integers

In scientific computing, most operations are on real numbers. Computations on integers rarely add up to any serious computation load[1]. It is mostly for completeness that we start with a short discussion of integers.

Integers are commonly stored in 16, 32, or 64 bits, with 16 becoming less common and 64 becoming more and more so. The main reason for this increase is not the changing nature of computations, but the fact that integers are used to index arrays. As the size of data sets grows (in particular in parallel computations), larger indices are needed. For instance, in 32 bits one can store the numbers zero through $2^{32} - 1 \approx 4 \cdot 10^9$. In other words, a 32 bit index can address 4 gigabytes of memory. Until recently this was enough for most purposes; these days the need for larger data sets has made 64 bit indexing necessary.

When we are indexing an array, only positive integers are needed. In general integer computations, of course, we need to accomodate the negative integers too. There are several ways of implementing negative integers. The

1. Some computations are done on bit strings. We will not mention them at all.

simplest solution is to reserve one bit as a *sign bit*, and use the remaining 31 (or 15 or 63; from now on we will consider 32 bits the standard) bits to store the absolute magnitude.

bitstring	$00\cdots0$	$\ldots$	$01\cdots1$	$10\cdots0$	$\ldots$	$11\cdots1$
interpretation as bitstring	0	$\ldots$	$2^{31}-1$	2^{31}	$\ldots$	$2^{32}-1$
interpretation as integer	0	$\cdots$	$2^{31}-1$	-0	$\cdots$	-2^{31}

This scheme has some disadvantages, one being that there is both a positive and negative number zero. This means that a test for equality becomes more complicated than simply testing for equality as a bitstring.

The scheme that is used most commonly is called *2's complement*, where integers are represented as follows.

- If $0 \leq m \leq 2^{31}-1$, the normal bit pattern for m is used.
- If $1 \leq n \leq 2^{31}$, then $-n$ is represented by the bit pattern for $2^{32}-n$.

bitstring	$00\cdots0$	$\ldots$	$01\cdots1$	$10\cdots0$	$\ldots$	$11\cdots1$
interpretation as bitstring	0	$\ldots$	$2^{31}-1$	2^{31}	$\ldots$	$2^{32}-1$
interpretation as integer	0	$\cdots$	$2^{31}-1$	$2^{32}-2^{31}$	$\cdots$	-1

Some observations:

- There is no overlap between the bit patterns for positive and negative integers, in particular, there is only one pattern for zero.
- The positive numbers have a leading bit zero, the negative numbers have the leading bit set.

Exercise 3.1. For the 'naive' scheme and the 2's complement scheme for negative numbers, give pseudocode for the comparison test $m < n$, where m and n are integers. Be careful to distinguish between all cases of m, n positive, zero, or negative.

Adding two numbers with the same sign, or multiplying two numbers of any sign, may lead to a result that is too large or too small to represent. This is called *overflow*.

Exercise 3.2. Investigate what happens when you perform such a calculation. What does your compiler say if you try to write down a nonrepresentible number explicitly, for instance in an assignment statement?

In exercise 1 above you explored comparing two integers. Let us know explore how subtracting numbers in two's complement is implemented. Consider $0 \leq m \leq 2^{31}-1$ and $1 \leq n \leq 2^{31}$ and let us see what happens in the computation of $m-n$.

Suppose we have an algorithm for adding and subtracting unsigned 32-bit numbers. Can we use that to subtract two's complement integers? We start by observing that the integer subtraction $m-n$ becomes the unsigned addition $m+(2^{32}-n)$.

- Case: $m < n$. Since $m + (2^{32} - n) = 2^{32} - (n - m)$, and $1 \leq n - m \leq 2^{31}$, we conclude that $2^{32} - (n - m)$ is a valid bit pattern. Moreover, it is the bit pattern representation of the negative number $m - n$, so we can indeed compute $m - n$ as an unsigned operation on the bitstring representations of m and n.
- Case: $m > n$. Here we observe that $m + (2^{32} - n) = 2^{32} + m - n$. Since $m - n > 0$, this is a number $> 2^{32}$ and therefore not a legitimate representation of a negative number. However, if we store this number in 33 bits, we see that it is the correct result $m - n$, plus a single bit in the 33-rd position. Thus, by performing the unsigned addition, and ignoring the *overflow bit*, we again get the correct result.

In both cases we conclude that we can perform subtraction by adding the bitstrings that represent the positive and negative number as unsigned integers, and ignoring overflow if it occurs.

3.2 Representation of real numbers

In this section we will look at how various kinds of numbers are represented in a computer, and the limitations of various schemes. The next section will then explore the ramifications of this on arithmetic involving computer numbers.

Real numbers are stored using a scheme that is analogous to what is known as 'scientific notation', where a number is represented as a *significant* and an *exponent*, for instance $6.022 \cdot 10^{23}$, which has a significant `6022` with a *radix point* after the first digit, and an exponent `23`. This number stands for

$$6.022 \cdot 10^{23} = \left[6 \times 10^0 + 0 \times 10^{-2} + 2 \times 10^{-2} + 2 \times 10^{-3}\right] \cdot 10^{24}.$$

We introduce a *base*, a small integer number, 10 in the preceding example, and 2 in computer numbers, and write numbers in terms of it as a sum of t terms:

$$x = \pm 1 \times \left[d_1\beta^0 + d_2\beta^{-1} + d_3\beta^{-2} + \cdots\right] \times \beta^e = \pm\Sigma_{i=1}^{t} d_i\beta^{1-i} \times \beta^e \tag{3.1}$$

where the components are

- the *sign bit*: a single bit storing whether the number is positive or negative;
- β is the base of the number system;
- $0 \leq d_i \leq \beta - 1$ the digits of the *mantissa* or *significant* – the location of the radix point (decimal point in decimal numbers) is implicitly assumed to the immediately following the first digit;
- t is the length of the mantissa;
- $e \in [L, U]$ exponent; typically $L < 0 < U$ and $L \approx -U$.

Note that there is an explicit sign bit for the whole number; the sign of the exponent is handled differently. For reasons of efficiency, e is not a signed number; instead it is considered as an unsigned number in *excess* of a certain minimum value. For instance, the bit pattern for the number zero is interpreted as $e = L$.

3.2.1 Some examples

Let us look at some specific examples of floating point representations. Base 10 is the most logical choice for human consumption, but computers are binary, so base 2 predominates there. Old IBM mainframes grouped bits to make for a base 16 representation.

	β	t	L	U
IEEE single precision (32 bit)	2	24	-126	127
IEEE double precision (64 bit)	2	53	-1022	1023
Old Cray 64 bit	2	48	-16383	16384
IBM mainframe 32 bit	16	6	-64	63
packed decimal	10	50	-999	999
Setun	3			

Of these, the single and double precision formats are by far the most common. We will discuss these in section 3.2.4 and further.

3.2.1.1 Binary coded decimal

Decimal numbers are not relevant in scientific computing, but they are useful in financial calculations, where computations involving money absolutely have to be exact. Binary arithmetic is at a disadvantage here, since numbers such as $1/10$ are repeating fractions in binary. With a finite number of bits in the mantissa, this means that the number $1/10$ can not be represented exactly in binary. For this reason, *binary-coded-decimal* schemes were used in old IBM mainframes, and are in fact being standardized in revisions of IEEE754 [3]; see also section 3.2.4. Few processors these days have hardware support for BCD; one example is the IBM Power6.

In BCD schemes, one or more decimal digits are encoded in a number of bits. The simplest scheme would encode the digits $0 \ldots 9$ in four bits. This has the advantage that in a BCD number each digit is readily identified; it has the disadvantage that about $1/3$ of all bits are wasted, since 4 bits can encode the numbers $0 \ldots 15$. More efficient encodings would encode $0 \ldots 999$ in ten bits, which could in principle store the numbers $0 \ldots 1023$. While this is efficient in the sense that few bits are wasted, identifying individual digits in such a number takes some decoding.

3.2.1.2 Ternary computers

There have been some experiments with ternary arithmetic [2, 7, 8].

3.2.2 Limitations

Since we use only a finite number of bits to store floating point numbers, not all numbers can be represented. The ones that can not be represented fall into two categories: those that are too large or too small (in some sense), and those that fall in the gaps. Numbers can be too large or too small in the following ways.

Overflow The largest number we can store is $(1 - \beta^{-t-1})\beta^U$, and the smallest number (in an absolute sense) is $-(1 - \beta^{-t-1})\beta^U$; anything larger than the former or smaller than the latter causes a condition called *overflow*.

Underflow The number closest to zero is $\beta^{-t-1} \cdot \beta^L$. A computation that has a result less than that (in absolute value) causes a condition called *underflow*. In fact, most computers use *normalized floating point numbers*: the first digit d_1 is taken to be nonzero; see section 3.2.3 for more about this. In this case, any number less than $\beta^{-1} \cdot \beta^L$ causes underflow. Trying to compute a number less than that is sometimes handled by using *unnormalized floating point numbers* (a process known as *gradual underflow*), but this is typically tens or hundreds of times slower than computing with regular floating point numbers. At the time of this writing, only the IBM Power6 has hardware support for gradual underflow.

The fact that only a small number of real numbers can be represented exactly is the basis of the field of round-off error analysis. We will study this in some detail in the following sections.

For detailed discussions, see the book by Overton [68]; it is easy to find online copies of the essay by Goldberg [41]. For extensive discussions of round-off error analysis in algorithms, see the books by Higham [51] and Wilkinson [86].

3.2.3 Normalized numbers and machine precision

The general definition of floating point numbers, equation (3.1), leaves us with the problem that numbers have more than one representation. For instance, $.5 \times 10^2 = .05 \times 10^3$. Since this would make computer arithmetic needlessly complicated, for instance in testing equality of numbers, we use *normalized floating point numbers*. A number is normalized if its first digit is nonzero. The implies that the mantissa part is $\beta > x_m \geq 1$.

A practical implication in the case of binary numbers is that the first digit is always 1, so we do not need to store it explicitly. In the IEEE 754 standard, this means that every floating point number is of the form

$$1.d_1 d_2 \ldots d_t \times 2^{exp}.$$

We can now be a bit more precise about the *representation error*. A machine number $\tilde{x}$ is the representation for all x in an interval around it. With t digits in the mantissa, this is the interval of numbers that differ from $\bar{x}$ in the $t+1$st digit. For the mantissa part we get:

$$\begin{cases} x \in \left[\tilde{x}, \tilde{x} + \beta^{-t}\right) & \text{truncation} \\ x \in \left[\tilde{x} - \frac{1}{2}\beta^{-t}, \tilde{x} + \frac{1}{2}\beta^{-t}\right) & \text{rounding} \end{cases}$$

Often we are only interested in the order of magnitude of the error, and we will write $\tilde{x} = x(1+\epsilon)$, where $|\epsilon| \leq \beta^{-t}$. This maximum relative error is called the *machine precision*, or sometimes *machine epsilon*. Typical values are:

$$\begin{cases} \epsilon \approx 10^{-7} & \text{32-bit single precision} \\ \epsilon \approx 10^{-16} & \text{64-bit double precision} \end{cases}$$

Machine precision can be defined another way: ϵ is the smallest number that can be added to 1 so that $1 + \epsilon$ has a different representation than 1. A small example shows how aligning exponents can shift a too small operand so

sign	exponent	mantissa
s	$e_1 \cdots e_8$	$s_1 \ldots s_{23}$
31	$30 \cdots 23$	$22 \cdots 0$

$(e_1 \cdots e_8)$	numerical value
$(0 \cdots 0) = 0$	$\pm 0.s_1 \cdots s_{23} \times 2^{-126}$
$(0 \cdots 01) = 1$	$\pm 1.s_1 \cdots s_{23} \times 2^{-126}$
$(0 \cdots 010) = 2$	$\pm 1.s_1 \cdots s_{23} \times 2^{-125}$
$\cdots$	
$(01111111) = 127$	$\pm 1.s_1 \cdots s_{23} \times 2^{0}$
$(10000000) = 128$	$\pm 1.s_1 \cdots s_{23} \times 2^{1}$
$\cdots$	
$(11111110) = 254$	$\pm 1.s_1 \cdots s_{23} \times 2^{127}$
$(11111111) = 255$	$\pm\infty$ if $s_1 \cdots s_{23} = 0$, otherwise `NaN`

Figure 3.1: Single precision arithmetic

that it is effectively ignored in the addition operation:

$$\begin{array}{rll} & 1.0000 & \times 10^0 \\ + & 1.0000 & \times 10^{-5} \\ \hline \end{array} \quad \Rightarrow \quad \begin{array}{rll} & 1.0000 & \times 10^0 \\ + & 0.00001 & \times 10^0 \\ \hline = & 1.0000 & \times 10^0 \end{array}$$

Yet another way of looking at this is to observe that, in the addition $x + y$, if the ratio of x and y is too large, the result will be identical to x.

The machine precision is the maximum attainable accuracy of computations: it does not make sense to ask for more than 6-or-so digits accuracy in single precision, or 15 in double.

Exercise 3.3. Write a small program that computes the machine epsilon. Does it make any difference if you set the compiler optimization levels low or high? Can you find other ways in which this computation goes wrong?

3.2.4 The IEEE 754 standard for floating point numbers

Some decades ago, issues like the length of the mantissa and the rounding behaviour of operations could differ between computer manufacturers, and even between models from one manufacturer. This was obviously a bad situation from a point of portability of codes and reproducibility of results. The IEE standard 754[2] codified all this, for instance stipulating 24 and 53 bits for the mantissa in single and double precision arithmetic, using a storage sequence of sign bit, exponent, mantissa. This for instance facilitates comparison of numbers[3].

2. IEEE 754 is a standard for binary arithmetic; there is a further standard, IEEE 854, that allows decimal arithmetic.

3. Computer systems can still differ as to how to store successive bytes. If the *least significant byte* is stored first, the system is called *little-endian*; if the *most significant byte* is stored first, it is called *big-endian*. See `http://en.wikipedia.org/wiki/Endianness` for details.

The standard also declared the rounding behaviour to be 'exact rounding': the result of an operation should be the rounded version of the exact result.

Above (section 3.2.2), we have seen the phenomena of overflow and underflow, that is, operations leading to unrepresentible numbers. There is a further exceptional situation that needs to be dealt with: what result should be returned if the program asks for illegal operations such as $\sqrt{-4}$? The IEEE 754 standard has two special quantities for this: `Inf` and `NaN` for 'infinity' and 'not a number'. If `NaN` appears in an expression, the whole expression will evaluate to that value. The rule for computing with `Inf` is a bit more complicated [41].

An inventory of the meaning of all bit patterns in IEEE 754 double precision is given in figure 3.1. Note that for normalized numbers the first nonzero digit is a 1, which is not stored, so the bit pattern $d_1 d_2 \ldots d_t$ is interpreted as $1.d_1 d_2 \ldots d_t$.

These days, almost all processors adhere to the IEEE 754 standard, with only occasional exceptions. For instance, Nvidia Tesla *GPU*s are not standard-conforming in single precision; see `http://en.wikipedia.org/wiki/Nvidia_Tesla`. The justification for this is that double precision is the 'scientific' mode, while sigle precision is mostly likely used for graphics, where exact compliance matters less.

3.3 Round-off error analysis

The fact that floating point numbers can only represent a small fraction of all real numbers, means that in practical circumstances a computation will hardly ever be exact. In this section we will study the phenomenon that most real numbers can not be represented, and what it means for the accuracy of computations. This is commonly called *round-off error analysis*.

3.3.1 Representation error

Numbers that are too large or too small to be represented are uncommon: usually computations can be arranged so that this situation will not occur. By contrast, the case that the result of a computation between computer numbers (even something as simple as a single addition) is not representable is very common. Thus, looking at the implementation of an algorithm, we need to analyze the effect of such small errors propagating through the computation. We start by analyzing the error between numbers that can be represented exactly, and those close by that can not be.

If x is a number and $\tilde{x}$ its representation in the computer, we call $x - \tilde{x}$ the *representation error* or *absolute representation error*, and $\frac{x-\tilde{x}}{x}$ the *relative representation error*. Often we are not interested in the sign of the error, so we may apply the terms error and relative error to $|x - \tilde{x}|$ and $|\frac{x-\tilde{x}}{x}|$ respectively.

Often we are only interested in bounds on the error. If ϵ is a bound on the error, we will write

$$\tilde{x} = x \pm \epsilon \underset{\text{D}}{\equiv} |x - \tilde{x}| \leq \epsilon \Leftrightarrow \tilde{x} \in [x - \epsilon, x + \epsilon]$$

For the relative error we note that

$$\tilde{x} = x(1 + \epsilon) \Leftrightarrow \left| \frac{\tilde{x} - x}{x} \right| \leq \epsilon$$

Let us consider an example in decimal arithmetic, that is, $\beta = 10$, and with a 3-digit mantissa: $t = 3$. The number $x = .1256$ has a representation that depends on whether we round or truncate: $\tilde{x}_{\mathrm{round}} = .126$, $\tilde{x}_{\mathrm{truncate}} = .125$. The error is in the 4th digit: if $\epsilon = x - \tilde{x}$ then $|\epsilon| < \beta^t$.

Exercise 3.4. The number in this example had no exponent part. What are the error and relative error if there had been one?

3.3.2 Correct rounding

The IEEE 754 standard, mentioned in section 3.2.4, does not only declare the way a floating point number is stored, it also gives a standard for the accuracy of operations such as addition, subtraction, multiplication, division. The model for arithmetic in the standard is that of *correct rounding*: the result of an operation should be as if the following procedure is followed:

- The exact result of the operation is computed, whether this is representable or not;
- This result is then rounded to the nearest computer number.

In short: the representation of the result of an operation is the rounded exact result of that operation. (Of course, after two operations it no longer needs to hold that the computed result is the exact rounded version of the exact result.)

If this statement sounds trivial or self-evident, consider subtraction as an example. In a decimal number system with two digits in the mantissa, the computation $.10 - .94 \cdot 10^{-1} = .10 - .094 = .006 = .06 \cdot 10^{-2}$. Note that in an intermediate step the mantissa $.094$ appears, which has one more digit than the two we declared for our number system. The extra digit is called a *guard digit*.

Without a guard digit, this operation would have proceeded as $.10 - .94 \cdot 10^{-1}$, where $.94 \cdot 10^{-1}$ would be rounded to $.09$, giving a final result of $.01$, which is almost double the correct result.

Exercise 3.5. Consider the computation $.10 - .95 \cdot 10^{-1}$, and assume again that numbers are rounded to fit the 2-digit mantissa. Why is this computation in a way a lot worse than the example?

One guard digit is not enough to guarantee correct rounding. An analysis that we will not reproduce here shows that three extra bits are needed.

3.3.3 Addition

Addition of two floating point numbers is done in a couple of steps. First the exponents are aligned: the smaller of the two numbers is written to have the same exponent as the larger number. Then the mantissas are added. Finally, the result is adjusted so that it again is a normalized number.

As an example, consider $.100 + .200 \times 10^{-2}$. Aligning the exponents, this becomes $.100 + .002 = .102$, and this result requires no final adjustment. We note that this computation was exact, but the sum $.100 + .255 \times 10^{-2}$ has the same result, and here the computation is clearly not exact. The error is $|.10255 - .102| < 10^{-3}$, and we note that the mantissa has 3 digits, so there clearly is a relation with the machine precision here.

In the example $.615 \times 10^1 + .398 \times 10^1 = 1.013 \times 10^2 = .101 \times 10^1$ we see that after addition of the mantissas an adjustment of the exponent is needed. The error again comes from truncating or rounding the first digit of the result that does not fit in the mantissa: if x is the true sum and $\tilde{x}$ the computed sum, then $\tilde{x} = x(1+\epsilon)$ where, with a 3-digit mantissa $|\epsilon| < 10^{-3}$.

Formally, let us consider the computation of $s = x_1 + x_2$, and we assume that the numbers x_i are represented as $\tilde{x}_i = x_i(1+\epsilon_i)$. Then the sum s is represented as

$$\begin{aligned} \tilde{s} &= (\tilde{x}_1 + \tilde{x}_2)(1+\epsilon_3) \\ &= x_1(1+\epsilon_1)(1+\epsilon_3) + x_2(1+\epsilon_2)(1+\epsilon_3) \\ &\approx x_1(1+\epsilon_1+\epsilon_3) + x_2(1+\epsilon_1+\epsilon_3) \\ &\approx s(1+2\epsilon) \end{aligned}$$

under the assumptions that all ϵ_i are small and of roughly equal size, and that both $x_i > 0$. We see that the relative errors are added under addition.

3.3.4 Multiplication

Floating point multiplication, like addition, involves several steps. In order to multiply two numbers $.m_1 \times \beta^{e_1}$ and $.m_2 \times \beta^{e_2}$, the following steps are needed.

- The exponents are added: $e \leftarrow e_1 + e_2$.
- The mantissas are multiplied: $m \leftarrow m_1 \times m_2$.
- The mantissa is normalized, and the exponent adjusted accordingly.

For example: $.123 \cdot 10^0 \times .567 \cdot 10^1 = .069741 \cdot 10^1 \rightarrow .69741 \cdot 10^0 \rightarrow .697 \cdot 10^0$.

What happens with relative errors?

3.3.5 Subtraction

Subtraction behaves very differently from addition. Whereas in addition errors are added, giving only a gradual increase of overall roundoff error, subtraction has the potential for greatly increased error in a single operation.

For example, consider subtraction with 3 digits to the mantissa: $.124 - .123 = .001 \rightarrow .100 \cdot 10^{-2}$. While the result is exact, it has only one significant digit[4]. To see this, consider the case where the first operand $.124$ is actually the rounded result of a computation that should have resulted in $.1235$. In that case, the result of the subtraction should have been $.050 \cdot 10^{-2}$, that is, there is a 100% error, even though the relative error of the inputs was as small as could be expected. Clearly, subsequent operations involving the result of this subtraction will also be inaccurate. We conclude that subtracting almost equal numbers is a likely cause of numerical roundoff.

There are some subtleties about this example. Subtraction of almost equal numbers is exact, and we have the correct rounding behaviour of IEEE arithmetic. Still, the correctness of a single operation does not imply that a sequence

4. Normally, a number with 3 digits to the mantissa suggests an error corresponding to rounding or truncating the fourth digit. We say that such a number has 3 *significant digits*. In this case, the last two digits have no meaning, resulting from the normalization process.

of operations containing it will be accurate. While the addition example showed only modest decrease of numerical accuracy, the cancellation in this example can have disastrous effects.

3.3.6 Examples

From the above, the reader may got the impression that roundoff errors only lead to serious problems in exceptional circumstances. In this section we will discuss some very practical examples where the inexactness of computer arithmetic becomes visible in the result of a computation. These will be fairly simple examples; more complicated examples exist that are outside the scope of this book, such as the instability of matrix inversion. The interested reader is referred to [86, 51].

3.3.6.1 The 'abc-formula'

As a practical example, consider the quadratic equation $ax^2 + bx + c = 0$ which has solutions $x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$. Suppose $b > 0$ and $b^2 \gg 4ac$ then $\sqrt{b^2 - 4ac} \approx b$ and the '$+$' solution will be inaccurate. In this case it is better to compute $x_- = \frac{-b-\sqrt{b^2-4ac}}{2a}$ and use $x_+ \cdot x_- = -c/a$.

3.3.6.2 Summing series

The previous example was about preventing a large roundoff error in a single operation. This example shows that even gradual buildup of roundoff error can be handled in different ways.

Consider the sum $\Sigma_{n=1}^{10000} \frac{1}{n^2} = 1.644834$ and assume we are working with single precision, which on most computers means a machine precision of 10^{-7}. The problem with this example is that both the ratio between terms, and the ratio of terms to partial sums, is ever increasing. In section 3.2.3 we observed that a too large ratio can lead to one operand of an addition in effect being ignored.

If we sum the series in the sequence it is given, we observe that the first term is 1, so all partial sums (Σ_n^N where $N < 10000$) are at least 1. This means that any term where $1/n^2 < 10^{-7}$ gets ignored since it is less than the machine precision. Specifically, the last 7000 terms are ignored, and the computed sum is 1.644725. The first 4 digits are correct.

However, if we evaluate the sum in reverse order we obtain the exact result in single precision. We are still adding small quantities to larger ones, but now the ratio will never be as bad as one-to-ϵ, so the smaller number is never ignored. To see this, consider the ratio of two terms subsequent terms:

$$\frac{n^2}{(n-1)^2} = \frac{n^2}{n^2 - 2n + 1} = \frac{1}{1 - 2/n + 1/n^2} \approx 1 + \frac{2}{n}$$

Since we only sum 10^5 terms and the machine precision is 10^{-7}, in the addition $1/n^2 + 1/(n-1)^2$ the second term will not be wholly ignored as it is when we sum from large to small.

Exercise 3.6. There is still a step missing in our reasoning. We have shown that in adding two subsequent terms, the smaller one is not ignored. However, during the calculation we add partial sums to the next term in the sequence. Show that this does not worsen the situation.

The lesson here is that series that are monotone (or close to monotone) should be summed from small to large, since the error is minimized if the quantities to be added are closer in magnitude. Note that this is the opposite strategy from the case of subtraction, where operations involving similar quantities lead to larger errors. This implies that if an application asks for adding and subtracting series of numbers, and we know a priori which terms are positive and negative, it may pay off to rearrange the algorithm accordingly.

3.3.6.3 Unstable algorithms

We will now consider an example where we can give a direct argument that the algorithm can not cope with problems due to inexactly represented real numbers.

Consider the recurrence $y_n = \int_0^1 \frac{x^n}{x-5}dx = \frac{1}{n} - 5y_{n-1}$.This is easily seen to be monotonically decreasing; the first term can be computed as $y_0 = \ln 6 - \ln 5$.

Performing the computation in 3 decimal digits we get:

computation		correct result
$y_0 = \ln 6 - \ln 5 = .182\|322 \times 10^1 \ldots$		1.82
$y_1 = .900 \times 10^{-1}$		.884
$y_2 = .500 \times 10^{-1}$		.0580
$y_3 = .830 \times 10^{-1}$	going up?	.0431
$y_4 = -.165$	negative?	.0343

We see that the computed results are quickly not just inaccurate, but actually nonsensical. We can analyze why this is the case.

If we define the error ϵ_n in the n-th step as

$$\tilde{y}_n - y_n = \epsilon_n,$$

then

$$\tilde{y}_n = 1/n - 5\tilde{y}_{n-1} = 1/n + 5n_{n-1} + 5\epsilon_{n-1} = y_n + 5\epsilon_{n-1}$$

so $\epsilon_n \geq 5\epsilon_{n-1}$. The error made by this computation shows exponential growth.

3.3.6.4 Linear system solving

Sometimes we can make statements about the numerical precision of a problem even without specifying what algorithm we use. Suppose we want to solve a linear system, that is, we have an $n \times n$ matrix A and a vector b of size n, and we want to compute the vector x such that $Ax = b$. (We will actually considering algorithms for this in chapter 5.) Since the vector b will the result of some computation or measurement, we are actually dealing with a vector $\tilde{b}$, which is some perturbation of the ideal b:

$$\tilde{b} = b + \Delta b.$$

The perturbation vector Δb can be of the order of the machine precision if it only arises from representation error, or it can be larger, depending on the calculations that produced $\tilde{b}$.

We now ask what the relation is between the exact value of x, which we would have obtained from doing an exact calculation with A and b, which is clearly impossible, and the computed value $\tilde{x}$, which we get from computing with A and $\tilde{b}$. (In this discussion we will assume that A itself is exact, but this is a simplification.)

Writing $\tilde{x} = x + \Delta x$, the result of our computation is now

$$A\tilde{x} = \tilde{b}$$

or

$$A(x + \Delta x) = b + \Delta b.$$

Since $Ax = b$, we get $A\Delta x = \Delta b$. From this, we get (see appendix A.1 for details)

$$\left\{ \begin{array}{ll} \Delta x & = A^{-1}\Delta b \\ Ax & = b \end{array} \right\} \Rightarrow \left\{ \begin{array}{ll} \|A\|\|x\| & \geq \|b\| \\ \|\Delta x\| & \leq \|A^{-1}\|\|\Delta b\| \end{array} \right. \Rightarrow \frac{\|\Delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\| \frac{\|\Delta b\|}{\|b\|} \tag{3.2}$$

The quantity $\|A\|\|A^{-1}\|$ is called the *condition number* of a matrix. The bound (3.2) then says that any perturbation in the right hand side can lead to a perturbation in the solution that is at most larger by the condition number of the matrix A. Note that it does not say that the perturbation in x *needs* to be anywhere close to that size, but we can not rule it out, and in some cases it indeed happens that this bound is attained.

Suppose that b is exact up to machine precision, and the condition number of A is 10^4. The bound (3.2) is often interpreted as saying that the last 4 digits of x are unreliable, or that the computation 'loses 4 digits of accuracy'.

3.3.7 Roundoff error in parallel computations

From the above example of summing a series we saw that addition in computer arithmetic is not associative. A similar fact holds for multiplication. This has an interesting consequence for parallel computations: the way a computation is spread over parallel processors influences the result. For instance, consider computing the sum of a large number N of terms. With P processors at our disposition, we can let each compute N/P terms, and combine the partial results. We immediately see that for no two values of P will the results be identical. This means that reproducibility of results in a parallel context is elusive.

3.4 More about floating point arithmetic

3.4.1 Programming languages

Different languages have different approaches to storing integers and floating point numbers.

Fortran In Fortran it is possible to specify the number of bytes that a number takes up: `INTEGER*2, REAL*8`. Often it is possible to write a code using only `INTEGER, REAL`, and use compiler flags to indicate the size of an integer and real number.

C In C, the type identifiers have no standard length. For integers there is `short int, int, long int`, and for floating point `float, double`. The `sizeof()` operator gives the number of bytes used to store a datatype.

C99, Fortran2003 Recent standards of the C and Fortran languages incorporate the C/Fortran interoperability standard, which can be used to declare a type in one language so that it is compatible with a certain type in the other language.

3.4.2 Other computer arithmetic systems

Other systems have been proposed to dealing with the problems of inexact arithmetic on computers. One solution is extended precision arithmetic, where numbers are stored in more bits than usual. A common use of this is in the calculation of inner products of vectors: the accumulation is internally performed in extended precision, but returned as a regular floating point number. Alternatively, there are libraries such as GMPlib [38] that allow for any calculation to be performed in higher precision.

Another solution to the imprecisions of computer arithmetic is 'interval arithmetic' [53], where for each calculation interval bounds are maintained. While this has been researched for considerable time, it is not practically used other than through specialized libraries [16].

3.4.3 Fixed-point arithmetic

A fixed-point number [87] can be represented as $\langle N, F\rangle$ where $N \geq \beta^0$ is the integer part and $F < 1$ is the fractional part. Another way of looking at this, is that a fixed-point number is an integer stored in $N + F$ digits, with an implied decimal point after the first N digits.

Fixed-point calculations can overflow, with no possibility to adjust an exponent. Consider the multiplication $\langle N_1, F_1\rangle \times \langle N_2, F_2\rangle$, where $N_1 \geq \beta^{n_1}$ and $N_2 \geq \beta^{n_2}$. This overflows if $n_1 + n_2$ is more than the number of positions available for the integer part. (Informally, the number of digits of the product is the sum of the digits of the operands.) This means that, in a program that uses fixed-point, numbers will need to have a number of zero digits, if you are ever going to multiply them, which lowers the numerical accuracy. It also means that the programmer has to think harder about calculations, arranging them in such a way that overflow will not occur, and that numerical accuracy is still preserved to a reasonable extent.

So why would people use fixed-point numbers? One important application is in embedded low-power devices, think a battery-powered digital thermometer. Since fixed-point calculations are essentially identical to integer calculations, they do not require a floating-point unit, thereby lowering chip size and lessening power demands. Also, many early video game systems had a processor that either had no floating-point unit, or where the integer unit was considerably faster than the floating-point unit. In both cases, implementing non-integer calculations as fixed-point, using the integer unit, was the key to high throughput.

Another area where fixed point arithmetic is still used, is in signal processing. In modern CPUs, integer and floating point operations are of essentially the same speed, but converting between them is relatively slow. Now, if the sine function is implemented through table lookup, this means that in $\sin(\sin x)$ the output of a function is used to index the next function application. Obviously, outputting the sine function in fixed point obviates the need for conversion between real and integer quantities, which simplifies the chip logic needed, and speeds up calculations.

3.4.4 Complex numbers

Some programming languages have complex numbers as a native data type, others not, and others are in between. For instance, in Fortran you can declare

```
COMPLEX z1,z2, z(32)
COMPLEX*16 zz1, zz2, zz(36)
```

A complex number is a pair of real numbers, the real and imaginary part, allocated adjacent in memory. The first declaration then uses 8 bytes to store to `REAL*4` numbers, the second one has `REAL*8`s for the real and imaginary part. (Alternatively, use `DOUBLE COMPLEX` or in Fortran90 `COMPLEX(KIND=2)` for the second line.)

By contrast, the `C` language does not natively have complex numbers, but both `C99` and `C++` have a `complex.h` header file[5]. This defines as complex number as in Fortran, as two real numbers.

Storing a complex number like this is easy, but sometimes it is computationally not the best solution. This becomes apparent when we look at arrays of complex numbers. If a computation often relies on access to the real (or imaginary) parts of complex numbers exclusively, striding through an array of complex numbers, has a stride two, which is disadvantageous (see section 1.2.4.3). In this case, it is better to allocate one array for the real parts, and another for the imaginary parts.

Exercise 3.7. Suppose arrays of complex numbers are stored the Fortran way. Analyze the memory access pattern of pairwise multiplying the arrays, that is, $\forall_i\colon c_i \leftarrow a_i \cdot b_i$, where `a(), b(), c()` are arrays of complex numbers.

Exercise 3.8. Show that an $n \times n$ linear system $Ax = b$ over the complex numbers can be written as a $2n \times 2n$ system over the real numbers. Hint: split the matrix and the vectors in their real and imaginary parts. Argue for the efficiency of storing arrays of complex numbers as separate arrays for the real and imaginary parts.

3.5 Conclusions

In a way, the reason for the error is the imperfection of computer arithmetic: if we could calculate with actual real numbers there would be no problem. However, if we accept roundoff as a fact of life, then various observations hold:

- Operations with 'the same' outcomes do not behave identically from a point of stability; see the 'abc-formula' example.

5. These two header files are not identical, and in fact not compatible. Beware, if you compile C code with a C++ compiler [29].

- Even rearrangements of the same computations do not behave identically; see the summing example.

Thus it becomes imperative to analyze computer algorithms with regard to their roundoff behaviour: does roundoff increase as a slowly growing function of problem parameters, such as the number of terms evaluted, or is worse behaviour possible? We will not address such questions in further detail in this book.

Chapter 4

Numerical treatment of differential equations

In this chapter we will look at the numerical solution of Ordinary Diffential Equations (ODEs) and Partial Diffential Equations (PDEs). These equations are commonly used in physics to describe phenomena such as the flow of air around an aircraft, or the bending of a bridge under various stresses. While these equations are often fairly simple, getting specific numbers out of them ('how much does this bridge sag if there are a hundred cars on it') is more complicated, often taking large computers to produce the desired results. Here we will describe the techniques that turn ODEs and PDEs into computable problems.

First of all, we will look at Initial Value Problems (IVPs), which describes processes that develop in time. Here we only consider ODEs: scalar functions that are only depend on time. The name derives from the fact that typically the function is specified at an initial time point.

Next, we will look at Boundary Value Problems (BVPs), describing processes in space. In realistic situations, this will concern multiple space variables, so we have a PDE. The name BVP is explained by the fact that the solution is specified on the boundary of the domain of definition.

Finally, we will consider the 'heat equation', an Initial Boundary Value Problems (IBVPs) which has aspects of both IVPs and BVPs: it describes heat spreading through a physical object such as a rod. The initial value describes the initial temperature, and the boundary values give prescribed temperatures at the ends of the rod.

For ease of analysis we will assume that all functions involved have sufficiently many higher derivatives, and that each derivative is sufficiently smooth.

4.1 Initial value problems

Many physical phenomena change over time, and typically the laws of physics give a description of the change, rather than of the quantity of interest itself. For instance, Newton's second law

$$F = ma$$

is a statement about the change in position of a point mass: expressed as

$$a = \frac{d^2}{dt^2}x = F/m$$

it states that acceleration depends linearly on the force exerted on the mass. A closed form description $x(t) = \ldots$ for the location of the mass can sometimes be derived analytically, but in many cases some form of approximation or numerical computation is needed.

Newton's equation is an ODE since it describes a function of one variable, time. It is an IVP since it describes the development in time, starting with some initial conditions. As an ODE, it is 'of second order' since it involves a second derivative, We can reduce this to first order, involving only first derivaties, if we allow vector quantities. Defining $u(t) = (x(t), x'(t))$, we find for u:

$$u' = Au + B, \qquad A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ F/a \end{pmatrix}$$

For simplicity, in this course we will only consider scalar equations; our reference equation is then

$$u'(t) = f(t, u(t)), \qquad u(0) = u_0, \qquad t > 0, \tag{4.1}$$

and in this section we will consider numerical methods for its solution.

Typically, the initial value in some starting point (often chosen as $t = 0$) is given: $u(0) = u_0$ for some value u_0, and we are interested in the behaviour of u as $t \to \infty$. As an example, $f(x) = x$ gives the equation $u'(t) = u(t)$. This is a simple model for population growth: the equation states that the rate of growth is equal to the size of the population. The equation (4.1) can be solved analytically for some choices of f, but we will not consider this. Instead, we only consider the numerical solution and the accuracy of this process.

In a numerical method, we consider discrete size time steps to approximate the solution of the continuous time-dependent process. Since this introduces a certain amount of error, we will analyze the error introduced in each time step, and how this adds up to a global error. In some cases, the need to limit the global error will impose restrictions on the numerical scheme.

4.1.1 Error and stability

Since numerical computation will always involve the inaccuracies stemming from the use of machine arithmetic, we want to avoid the situation where a small perturbation in the initial value leads to large perturbations in the solution. Therefore, we will call a differential equation 'stable' if solutions corresponding to different initial values u_0 converge to one another as $t \to \infty$.

Let us limit ourselves to the so-called 'autonomous' ODE

$$u'(t) = f(u(t)) \tag{4.2}$$

in which the right hand side does not explicitly depend on t[1]. A sufficient criterium for stability is:

$$\frac{\partial}{\partial u} f(u) = \begin{cases} > 0 & \text{unstable} \\ = 0 & \text{neutrally stable} \\ < 0 & \text{stable} \end{cases}$$

Proof. If u^* is a zero of f, meaning $f(u^*) = 0$, then the constant function $u(t) \equiv u^*$ is a solution of $u' = f(u)$, a so-called 'equilibrium' solution. We will now consider how small perturbations from the equilibrium behave. Let u be a solution of the PDE, and write $u(t) = u^* + \eta(t)$, then we have

$$\begin{aligned} \eta' &= u' = f(u) = f(u^* + \eta) = f(u^*) + \eta f'(u^*) + O(\eta^2) \\ &= \eta f'(u^*) + O(\eta^2) \end{aligned}$$

Ignoring the second order terms, this has the solution

$$\eta(t) = e^{f'(x^*)t}$$

which means that the perturbation will damp out if $f'(x^*) < 0$.

We will often refer to the simple example $f(u) = -\lambda u$, with solution $u(t) = u_0 e^{-\lambda t}$. This problem is stable if $\lambda > 0$.

4.1.2 Finite difference approximation

In order to solve the problem numerically, we turn the continuous problem into a discrete one by looking at finite time/space steps. Assuming all functions are sufficiently smooth, a straightforward Taylor expansion[2] gives:

$$u(t + \Delta t) = u(t) + u'(t)\Delta t + u''(t)\frac{\Delta t^2}{2!} + u'''(t)\frac{\Delta t^3}{3!} + \cdots$$

This gives for u':

$$u'(t) = \frac{u(t + \Delta t) - u(t)}{\Delta t} + O(\Delta t^2) \tag{4.3}$$

We see that we can approximate a differential operator by a *finite difference*, with an error that is known in its order of magnitude..

Substituting this in $u' = f(t, u)$ gives

$$\frac{u(t + \Delta t) - u(t)}{\Delta t} = f(t, u(t)) + O(\Delta t^2)$$

1. Non-autonomous ODEs can be transformed to autonomous ones, so this is no limitation.
2. See appendix A.5 if you are unfamiliar with this.

or

$$u(t+\Delta t) = u(t) + \Delta t\, f(t, u(t)) + O(\Delta t^3)$$

We use this equation to derive a numerical scheme: with $t_0 = 0$, $t_{k+1} = t_k + \Delta t = \cdots = (k+1)\Delta t$, $u(t_k) = u_k$ we get a difference equation

$$u_{k+1} = u_k + \Delta t\, f(t_k, u_k).$$

This is known as the 'Explicit Euler' or 'Euler forward' method.

The process of going from a differential equation to a difference equation is often referred to as *discretization*, since we compute function values only in a discrete set of points. The values computed themselves are still real valued. Another way of phrasing this: the numerical solution is found in a finite dimensional, or countably infinite dimensional, space. The solution to the original problem is found in the space of functions $\mathbb{R} \to \mathbb{R}$.

In (4.3) we approximated one operator by another, and in doing so made a *truncation error* of $O(\Delta t^2)$. This does *not* immediately imply that the difference equation computes a solution that is close to the true solution. For that some more analysis is needed.

We start by analyzing the 'local error': if we assume the computed solution is exact at step k, that is, $u_k = u(t_k)$, how wrong will we be at step $k+1$? We have

$$\begin{aligned} u(t_{k+1}) &= u(t_k) + u'(t_k)\Delta t + u''(t_k)\tfrac{\Delta t^2}{2!} + \cdots \\ &= u(t_k) + f(t_k, u(t_k))\Delta t + u''(t_k)\tfrac{\Delta t^2}{2!} + \cdots \end{aligned}$$

and

$$u_{k+1} = u_k + f(t_k u_k)\Delta t$$

So

$$\begin{aligned} L_{k+1} &= u_{k+1} - u(t_{k+1}) = u_k - u(t_k) + f(t_k, u_k) - f(t_k, u(t_k)) - u''(t_k)\frac{\Delta t^2}{2!} + \cdots \\ &= -u''(t_k)\frac{\Delta t^2}{2!} + \cdots \end{aligned}$$

This shows that in each step we make an error of $O(\Delta t^2)$. If we assume that these errors can be added, we find a global error of

$$E_k \approx \Sigma_k L_k = k\Delta t \frac{\Delta t^2}{2!} = O(\Delta t)$$

Since the global error is of first order in Δt, we call this a 'first order method'. Note that this error, which measures the distance between the true and computed solutions, is of lower order than the truncation error, which is the error in approximating the operator.

4.1.2.1 An Euler forward example

Consider the IVP $u' = f(t,u)$ for $t \geq 0$, where $f(t,u) = -\lambda u$ and an initial value $u(0) = u_0$ is given. This has an exact solution of $u(t) = u_0 e^{-\lambda t}$. From the above discussion, we conclude that this problem is stable, meaning that small perturbations in the solution ultimately damp out, if $\lambda > 0$. We will now investigate the question of whether the numerical solution behaves the same way as the exact solution, that is, whether numerical solutions also converge to zero.

The Euler forward, or explicit Euler, scheme for this problem is

$$u_{k+1} = u_k - \Delta t \lambda u_k = (1 - \lambda \Delta t) u_k = (1 - \lambda \Delta t)^k u_0$$

For stability, we require that $u_k \to 0$ as $k \to \infty$. This is equivalent to

$$\begin{aligned} u_k \to \infty &\Leftrightarrow |1 - \lambda \Delta t| < 1 \\ &\Leftrightarrow -1 < 1 - \lambda \Delta t < 1 \\ &\Leftrightarrow -2 < -\lambda \Delta t < 0 \\ &\Leftrightarrow 0 < \lambda \Delta t < 2 \\ &\Leftrightarrow \Delta t < 2/\lambda \end{aligned}$$

We see that the stability of the numerical solution scheme depends on the value of Δt: the scheme is only stable if Δt is small enough. For this reason, we call the explicit Euler method *conditionally stable*. Note that the stability of the differential equation and the stability of the numerical scheme are two different questions. The continuous problem is stable if $\lambda > 0$; the numerical problem has an additional condition that depends on the discretization scheme used.

4.1.2.2 Implicit Euler

The explicit method you just saw was easy to compute, but the conditional stability is a potential problem. For instance, it could imply that the number of time steps would be a limiting factor. There is an alternative to the explicit method that does not suffer from the same objection.

Instead of expanding $u(t + \Delta t)$, consider the following:

$$u(t - \Delta t) = u(t) - u'(t)\Delta t + u''(t)\frac{\Delta t^2}{2!} + \cdots$$

which implies

$$u'(t) = \frac{u(t) - u(t - \Delta t)}{\Delta t} + u''(t)\Delta t/2 + \cdots$$

As before, we take the equation $u'(t) = f(t, u(t))$ and approximate $u'(t)$ by a difference formula:

$$\frac{u(t) - u(t - \Delta t)}{\Delta t} = f(t, u(t)) + O(\Delta t) \Rightarrow u(t) = u(t - \Delta t) + \Delta t f(t, u(t)) + O(\Delta t^2)$$

This inspires us to take fixed points $u_k \equiv u(kt)$ and define a numerical scheme:

$$u_{k+1} = u_k + \Delta t f(t_{k+1}, u_{k+1}).$$

An important difference with the explicit scheme is that u_{k+1} now also appears on the right hand side of the equation. That is, compution of u_{k+1} is now implicit. For example, let $f(t,u) = -u^3$, then $u_{k+1} = u_k - \Delta t u_{k+1}^3$. This needs a way to solve a nonlinear equation; typically this can be done with Newton iterations.

4.1.2.3 Stability of Implicit Euler

Let us take another look at the example $f(t,u) = -\lambda u$. Formulating the implicit method gives

$$u_{k+1} = u_k - \lambda \Delta t u_{k+1} \Leftrightarrow (1 + \Delta t) u_{k+1} = u_k$$

so

$$u_{k+1} = \left(\frac{1}{1+\lambda\Delta t}\right) u_k = \left(\frac{1}{1+\lambda\Delta t}\right)^k u_0$$

If $\lambda > 0$, which is the condition for a stable equation, we find that $u_k \to 0$ for all values of λ and Δt. This method is called *unconditionally stable*. The main advantage of an implicit method over an explicit one is clearly the stability: it is possible to take larger time steps without worrying about unphysical behaviour. Of course, large time steps can make convergence to the *steady state* (see Appendix A.4.4) slower, but at least there will be no divergence.

On the other hand, implicit methods are more complicated. As you saw above, they can involve nonlinear systems to be solved in every time step. In cases where u is vector-valued, such as in the heat equation, discussed below, you will see that the implicit method requires the solution of a system of equations.

Exercise 4.1. Analyse the accuracy and computational aspects of the following scheme for the IVP $u'(x) = f(x)$:

$$u_{i+1} = u_i + h(f(x_i) + f(x_{i+1}))/2$$

which corresponds to adding the Euler explicit and implicit schemes together. You do not have to analyze the stability of this scheme.

Exercise 4.2. Consider the initial value problem $y'(t) = y(t)(1 - y(t))$. Observe that $y \equiv 0$ and $y \equiv 1$ are solutions. These are called 'equilibrium solutions'.

1. A solution is stable, if perturbations 'converge back to the solution', meaning that for ϵ small enough,

   $$\text{if } y(t) = \epsilon \text{ for some } t, \text{ then } \lim_{t\to\infty} y(t) = 0$$

   and

   $$\text{if } y(t) = 1 + \epsilon \text{ for some } t, \text{ then } \lim_{t\to\infty} y(t) = 1$$

This requires for instance that

$$y(t) = \epsilon \Rightarrow y'(t) < 0.$$

Investigate this behaviour. Is zero a stable solution? Is one?

2. Formulate an explicit method for computing a numerical solution to the differential equation. Show that

$$y_k \in (0,1) \Rightarrow y_{k+1} > y_k, \qquad y_k > 1 \Rightarrow y_{k+1} < y_k$$

3. Write a small program to investigate the behaviour of the numerical solution under various choices for Δt. Include program listing and a couple of runs in your homework submission.
4. You see from running your program that the numerical solution can oscillate. Derive a condition on Δt that makes the numerical solution monotone. It is enough to show that $y_k < 1 \Rightarrow y_{k+1} < 1$, and $y_k > 1 \Rightarrow y_{k+1} > 1$.
5. Now formulate an implicit method, and show that y_{k+1} is easily computed from y_k. Write a program, and investigate the behaviour of the numerical solution under various choices for Δt.
6. Show that the numerical solution of the implicit scheme is monotone for all choices of Δt.

4.2 Boundary value problems

In the previous section you saw initial value problems, which model phenomena that evolve over time. We will now move on to'boundary value problems', which are in general stationary in time, but which describe a phenomenon that is location dependent. Examples would be the shape of a bridge under a load, or the heat distribution in a window pane, as the temperature outside differs from the one inside.

The general form of a (second order, one-dimensional) BVP is[3]

$$u''(x) = f(x, u, u') \text{ for } x \in [a,b] \text{ where } u(a) = u_a,\ u(b) = u_b$$

but here we will only consider the simple form

$$-u''(x) = f(x) \text{ for } x \in [0,1] \text{ with } u(0) = u_0,\ u(1) = u_1. \tag{4.4}$$

in one space dimension, or

$$-u_{xx}(\bar{x}) - u_{yy}(\bar{x}) = f(\bar{x}) \text{ for } \bar{x} \in \Omega = [0,1]^2 \text{ with } u(\bar{x}) = u_0 \text{ on } \delta\Omega. \tag{4.5}$$

in two space dimensions. Here, $\delta\Omega$ is the boundary of the domain Ω. Since we prescribe the value of u on the boundary, such a problem is called a Boundary Value Problem (BVP).

3. Actually, the boundary conditions are can be more general, involving derivatives on the interval end points.

4.2.1 General PDE theory

There are several types of PDE, each with distinct mathematical properties. The most important property is that of *region of influence*: if we tinker with the problem so that the solution changes in one point, what other points will be affected.

- Elliptic PDEs have the form

  $$Au_{xx} + Bu_{yy} + \text{lower order terms} = 0$$

  where $A, B > 0$. They are characterized by the fact that all points influence each other. These equations often describe phenomena in structural mechanics, such as a beam or a membrane. It is intuitively clear that pressing down on any point of a membrane will change the elevation of every other point, no matter how little.
- Hyperbolic PDEs are of the form

  $$Au_{xx} + Bu_{yy} + \text{lower order terms} = 0$$

  with A, B of opposite sign. Such equations describe wave phenomena. Intuitively, changing the solution at any point will only change certain future points, since waves have a propagation speed that makes it impossible for a point to influence points in the future that are too far away.
- Parabolic PDEs are of the form

  $$Au_x + Bu_{yy} + \text{no higher order terms in } x = 0$$

  and they describe diffusion-like phenomena. The best way to characterize them is to consider that the solution in each point in space and time is influenced by a certain finite region at each previous point in space[4].

4.2.2 The Poisson equation

We call the operator Δ, defined by

$$\Delta u = u_{xx} + u_{yy},$$

a second order *differential operator*, and equation (4.5) a second-order PDE. Specifically, the problem

$$-u_{xx}(\bar{x}) - u_{yy}(\bar{x}) = f(\bar{x}) \text{ for } \bar{x} \in \Omega = [0,1]^2 \text{ with } u(\bar{x}) = u_0 \text{ on } \delta\Omega. \tag{4.6}$$

is called the *Poisson equation*, in this case defined on the unit square. Second order PDEs are quite common, describing many phenomena in fluid and heat flow, and structural mechanics.

4. This leads to a condition limiting the time step in IBVPs, known as the Courant-Friedrichs-Lewy condition `http://en.wikipedia.org/wiki/CourantFriedrichsLewy_condition`. It describes the notion that in the exact problem $u(x,t)$ depends on a range of $u(x', t - \Delta t)$ values; the time step of the numerical method has to be small enough that the numerical solution takes all these points into account.

4.2.2.1 One-dimensional case

At first, for simplicity, we consider the one-dimensional Poisson equation

$$-u_{xx} = f(x).$$

In order to find a numerical scheme we use Taylor series as before, expressing $u(x+h)$ or $u(x-h)$ in terms of u and its derivatives at x. Let $h > 0$, then

$$u(x+h) = u(x) + u'(x)h + u''(x)\frac{h^2}{2!} + u'''(x)\frac{h^3}{3!} + u^{(4)}(x)\frac{h^4}{4!} + u^{(5)}(x)\frac{h^5}{5!} + \cdots$$

and

$$u(x-h) = u(x) - u'(x)h + u''(x)\frac{h^2}{2!} - u'''(x)\frac{h^3}{3!} + u^{(4)}(x)\frac{h^4}{4!} - u^{(5)}(x)\frac{h^5}{5!} + \cdots$$

Our aim is now to approximate $u''(x)$. We see that the u' terms in these equations would cancel out under addition, leaving $2u(x)$:

$$u(x+h) + u(x-h) = 2u(x) + u''(x)h^2 + u^{(4)}(x)\frac{h^4}{12} + \cdots$$

so

$$-u''(x) = \frac{2u(x) - u(x+h) - u(x-h)}{h^2} + u^{(4)}(x)\frac{h^2}{12} + \cdots \tag{4.7}$$

The basis for a numerical scheme for (4.4) is then the observation

$$\frac{2u(x) - u(x+h) - u(x-h)}{h^2} = f(x, u(x), u'(x)) + O(h^2),$$

which shows that we can approximate the differential operator by a difference operator, with an $O(h^2)$ *truncation error*.

To derive a numerical method, we divide the interval $[0,1]$ into equally spaced points: $x_k = kh$ where $h = 1/n$ and $k = 0 \ldots n$. With these, the Finite Difference (FD) formula (4.7) leads to a numerical scheme that forms a system of equations:

$$-u_{k+1} + 2u_k - u_{k-1} = h^2 f(x_k) \quad \text{for } k = 1, \ldots, n-1 \tag{4.8}$$

For most values of k this equation relates u_k unknown to the unknowns u_{k-1} and u_{k+1}. The exceptions are $k = 1$ and $k = n-1$. In that case we recall that u_0 and u_n are known boundary conditions, and we write the equations as

$$\begin{cases} 2u_1 - u_2 = h^2 f(x_1) + u_0 \\ 2u_{n-1} - u_{n-2} = h^2 f(x_{n-1}) + u_n \end{cases}$$

We can now summarize these equation for $u_k, k = 1 \ldots n-1$ as a matrix equation:

$$\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} h^2 f_1 + u_0 \\ h^2 f_2 \\ \vdots \end{pmatrix} \tag{4.9}$$

This has the form $Au = f$ with A a fully known matrix, f a fully known vector, and u a vector of unknowns. Note that the right hand side vector has the boundary values of the problem in the first and last locations. This means that, if you want to solve the same differential equation with different boundary conditions, only the vector f changes.

Exercise 4.3. A condition of the type $u(0) = u_0$ is called a *Dirichlet boundary condition*. Physically, this corresponds to knowing the temperature at the end point of a rod. Other boundary conditions exist. Specifying a value for the derivative, $u'(0) = u'_0$, rather than for the function value,would be appropriate if we are modeling fluid flow and the outflow rate at $x = 0$ is known. This is known as a *Neumann boundary condition*.
A Neumann boundary condition $u'(0) = u'_0$ can be modeled by stating

$$\frac{u_0 - u_1}{h} = u'_0.$$

Show that, unlike in the case of the Direchlet boundary condition, this affects the matrix of the linear system.
Show that having a Neumann boundary condition at both ends gives a singular matrix, and therefore no unique solution to the linear system. (Hint: guess the vector that has eigenvalue zero.)
Physically this makes sense. For instance, in an elasticity problem, Dirichlet boundary conditions state that the rod is clamped at a certain height; a Neumann boundary condition only states its angle at the end points, which leaves its height undetermined.

Let us list some properties of A that you will see later are relevant for solving such systems of equations:

- The matrix is very *sparse*: the percentage of elements that is nonzero is low. The nonzero elements are not randomly distributed but located in a band around the main diagonal. We call this a *band matrix* in general, and a *tridiagonal matrix* in this specific case.
- The matrix is symmetric. This property does not hold for all matrices that come from discretizing BVPs, but it is true if there are no odd order derivatives, such as u_x, u_{xxx}, u_{xy}.
- Matrix elements are constant in each diagonal, that is, in each set of points $\{(i,j)\colon i-j=c\}$ for some c. This is only true for very simple problems. It is no longer true if the differential equation has location dependent terms such as $\frac{d}{dx}(a(x)\frac{d}{dx}u(x))$. It is also no longer true if we make h variable through the interval, for instance because we want to model behaviour around the left end point in more detail.
- Matrix elements conform to the following sign pattern: the diagonal elements are positive, and the off-diagonal elements are nonpositive. This property depends on the numerical scheme used, but it is often true. Together with the following property of definiteness, this is called an *M-matrix*. There is a whole mathematical theory of these matrices [14].

- The matrix is positive definite: $x^t Ax > 0$ for all nonzero vectors x. This property is inherited from the original continuous problem, if the numerical scheme is carefully chosen. While the use of this may not seem clear at the moment, later you will see methods for solving the linear system that depend on it.

Strictly speaking the solution of equation (4.9) is simple: $u = A^{-1}f$. However, computing A^{-1} is not the best way of finding u. As observed just now, the matrix A has only $3N$ nonzero elements to store. Its inverse, on the other hand, does not have a single nonzero. Although we will not prove it, this sort of statement holds for most sparse matrices. Therefore, we want to solve $Au = f$ in a way that does not require $O(n^2)$ storage.

4.2.2.2 Two-dimensional BVPs

The one-dimensional BVP above was atypical in a number of ways, especially related to the resulting linear algebra problem. In this section we will see a two-dimensional problem, which displays some new aspects.

The one-dimensional problem above had a function $u = u(x)$, which now becomes $u = u(x, y)$ in two dimensions. The two-dimensional problem we are interested is then

$$-u_{xx} - u_{yy} = f, \quad (x, y) \in [0, 1]^2, \tag{4.10}$$

where the values on the boundaries are given. We get our discrete equation by applying equation (4.7) in x and y directions:

$$\begin{aligned} -u_{xx}(x, y) &= \tfrac{2u(x,y)-u(x+h,y)-u(x-h,y)}{h^2} + u^{(4)}(x, y)\tfrac{h^2}{12} + \cdots \\ -u_{yy}(x, y) &= \tfrac{2u(x,y)-u(x,y+h)-u(x,y-h)}{h^2} + u^{(4)}(x, y)\tfrac{h^2}{12} + \cdots \end{aligned}$$

or, taken together:

$$4u(x, y) - u(x + h, y) - u(x - h, y) - u(x, y + h) - u(x, y - h) = 1/h^2\, f(x, y) + O(h^2) \tag{4.11}$$

Let again $h = 1/n$ and define $x_i = ih$ and $y_j = jh$; let u_{ij} be the approximation to $u(x_i, y_j)$, then our discrete equation becomes

$$4u_{ij} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} = h^{-2} f_{ij}. \tag{4.12}$$

We now have $n \times n$ unknowns u_{ij} which we can put in a linear ordering by defining $I = I_{ij} = i + j \times n$. This gives us $N = n^2$ equations

$$4u_I - u_{I+1} - u_{I-1} - u_{I+n} - u_{I-n} = h^{-2} f_I. \tag{4.13}$$

Figure 4.1 shows how equation (4.12) connects domain points.

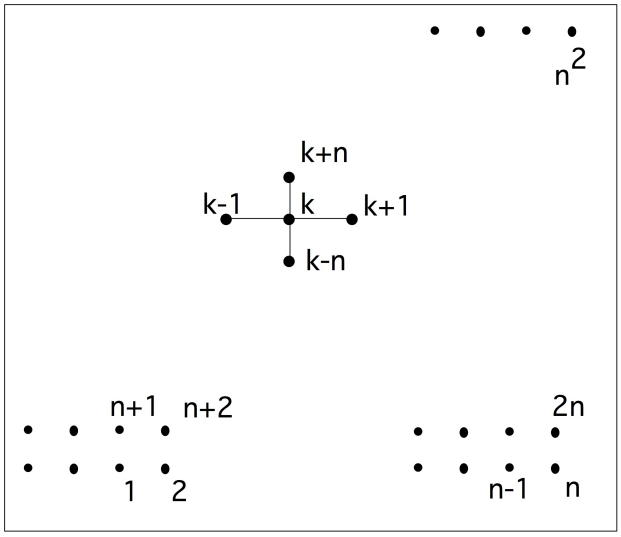


Figure 4.1: A difference stencil applied to a two-dimensional square domain

If we take the set of equation obtained by letting I in equation (4.13) range through the domain Ω, we find a linear system $Au = f$ of size N with a special structure:

$$A = \left(\begin{array}{cccccc|cccccc|c} 4 & -1 & & & \emptyset & & -1 & & & & \emptyset & & \\ -1 & 4 & 1 & & & & & -1 & & & & & \\ & \ddots & \ddots & \ddots & & & & & \ddots & & & & \\ & & \ddots & \ddots & -1 & & & & & \ddots & & & \\ \emptyset & & & -1 & 4 & & \emptyset & & & & -1 & & \\ \hline -1 & & & & \emptyset & & 4 & -1 & & & & & \\ & \ddots & & & & & -1 & \ddots & \ddots & & & \ddots & \\ \hline & & & & & & & & & & & \ddots & \end{array}\right) \tag{4.14}$$

The matrix is again banded, but unlike in the one-dimensional case, there are zeros inside the band. (This has some important consequences when we try to solve this system; see section 5.4.3.) Because the matrix has five nonzero diagonals, it is said to be of *penta-diagonal* structure.

You can also put a block structure on the matrix, by grouping the unknowns together that are in one row of the domain. This is called a *block matrix*, and, on the block level, it has a *tridiagonal* structure, so we call this a *block tridiagonal* matrix. Note that the diagonal blocks themselves are tridiagonal; the off-diagonal blocks are minus the identity matrix.

This matrix, like the one-dimensional example above, has constant diagonals, but this is again due to the simple nature of the problem. In practical problems it will not be true. That said, such 'constant coefficient' problems occur,

and when they are on rectangular domains, there are very efficient methods for solving linear system with $N \log N$ complexity.

Exercise 4.4. The block structure of the matrix, with all diagonal blocks having the same size, is due to the fact that we defined our BVP on a square domain. Sketch the matrix structure that arises from discretizing equation (4.10), again with central differences, but this time defined on a triangular domain. Show that, again, there is a block tridiagonal structure, but that the blocks are now of varying sizes.

For domains that are even more irregular, the matrix structure will also be irregular.

The regular block structure is also caused by our decision to order the unknowns by rows and columns. This known as the *natural ordering* or *lexicographic ordering*; various other orderings are possible. One common way of ordering the unknowns is the red-black ordering or checkerboard ordering which has advantanges for parallel computation. This will be discussed in section 6.7.2.

There is more to say about analytical aspects of the BVP (for instance, how smooth is the solution and how does that depend on the boundary conditions?) but those questions are outside the scope of this course. In the chapter on linear algebra, we will come back to the BVP, since solving the linear system is mathematically interesting.

4.2.2.3 Difference stencils

The discretization (4.11) is often phrased as applying the *difference stencil*

$$\begin{array}{ccc} \cdot & -1 & \cdot \\ -1 & 4 & -1 \\ \cdot & -1 & \cdot \end{array}$$

to the function u. Given a physical domain, we apply the stencil to each point in that domain to derive the equation for that point. Figure 4.1 illustrates that for a square domain of $n \times n$ points. Connecting this figure with equation (4.14), you see that the connections in the same line give rise to the main diagonal and first upper and lower offdiagonal; the connections to the next and previous lines become the nonzeros in the off-diagonal blocks.

This particular stencil is often referred to as the '5-point star'. There are other difference stencils; the structure of some of them are depicted in figure 4.2. A stencil with only connections in horizontal or vertical direction is called a 'star stencil', while one that has cross connections (such as the second in figure 4.2) is called a 'box stencil'.

Exercise 4.5. Consider the third stencil in figure 4.2, used for a BVP on a square domain. What does the sparsity structure of the resulting matrix look like, if we again order the variables by rows and columns?

Other stencils than the 5-point star can be used to attain higher accuracy, for instance giving a truncation error of $O(h^4)$. They can also be used for other differential equations than the one discussed above. For instance, it is not hard to show that the 5-point stencil can not give a discretization of the equation $u_{xxxx} + u_{yyyy} = f$ with less than $O(1)$ truncation error.

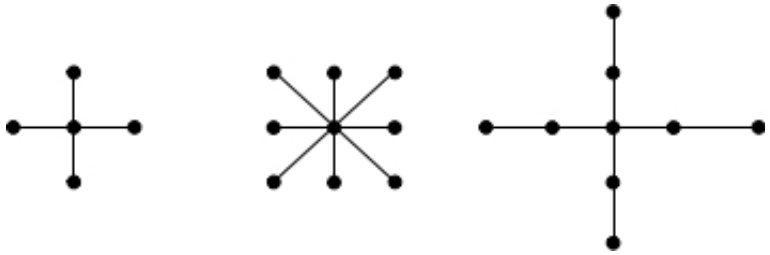


Figure 4.2: The structure of some difference stencils in two dimensions

While the discussion so far has been about two-dimensional problems, it is easily generalized to higher dimensions for such equations as $-u_{xx} - u_{yy} - u_{zz} = f$. The straightforward generalization of the 5-point stencil, for instance, becomes a 7-point stencil in three dimensions.

4.2.2.4 *Other discretization techniques*

In the above, we used finite differences to find a numerical solution to a differential equation. There are various other techniques, and in fact, in the case of boundary value problems, they are usually preferred over finite differences. The most popular methods are the *Finite Element Method (FEM)* and the *finite volume method*. Especially the finite element method is attractive, since it can handle irregular shapes more easily, and it is more amenable to approximation error analysis. However, on the simple problems discussed here, they give similar or even the same linear systems as FD methods.

There will be a brief discussion of finite element matrices in section 6.6.2.

4.3 Initial Boundary value problem

We will now go on to discuss an IBVP, which, as you may deduce from the name, combines aspects of IVPs and BVPs. Here we will limit ourselves to one space dimension.

The problem we are considering is that of heat conduction in a rod, where $T(x,t)$ describes the temperature in location x at time t, for $x \in [a,b]$, $t > 0$. The so-called *heat equation* (see Appendix A.4 for a quick introduction to PDEs in general and the heat equation in particular) is:

$$\frac{\partial}{\partial t}T(x,t) - \alpha\frac{\partial^2}{\partial x^2}T(x,t) = q(x,t)$$

where

- The initial condition $T(x,0) = T_0(x)$ describes the initial temperature distribution.
- The boundary conditions $T(a,t) = T_a(t)$, $T(b,t) = T_b(t)$ describe the ends of the rod, which can for instance be fixed to an object of a known temperature.

- The material the rod is made of is modeled by a single parameter $\alpha > 0$, the thermal diffusivity, which describes how fast heat diffuses through the material.
- The forcing function $q(x,t)$ describes externally applied heating, as a function of both time and place.

There is a simple connection between the IBVP and the BVP: if the boundary functions T_a and T_b are constant, and q does not depend on time, only on location, then intuitively T will converge to a *steady state*. The equation for this is $u''(x) = f$.

4.3.1 Discretization

We now discretize both space and time, by $x_{j+1} = x_j + \Delta x$ and $t_{k+1} = t_k + \Delta t$, with boundary conditions $x_0 = a$, $x_n = b$, and $t_0 = 0$. We write T_j^k for the numerical solution at $x = x_j, t = t_k$; with a little luck, this will approximate the exact solution $T(x_j, t_k)$.

For the space discretization we use the central difference formula (4.8):

$$\frac{\partial^2}{\partial x^2} T(x,t_k)\bigg|_{x=x_j} \Rightarrow \frac{T_{j-1}^k - 2T_j^k + T_{j+1}^k}{\Delta x^2}.$$

For the time discretization we can use any of the schemes in section 4.1.2. We will investigate again the explicit and implicit schemes, with similar conclusions about the resulting stability.

4.3.1.1 Explicit scheme

With explicit time stepping we approximat the time derivative as

$$\frac{\partial}{\partial t} T(x_j,t)\bigg|_{t=t_k} \Rightarrow \frac{T_j^{k+1} - T_j^k}{\Delta t}. \tag{4.15}$$

Taking this together with the central differences in space, we now have

$$\frac{T_j^{k+1} - T_j^k}{\Delta t} - \alpha \frac{T_{j-1}^k - 2T_j^k + T_{j+1}^k}{\Delta x^2} = q_j^k$$

which we rewrite as

$$T_j^{k+1} = T_j^k + \frac{\alpha \Delta t}{\Delta x^2}(T_{j-1}^k - 2T_j^k + T_{j+1}^k) + \Delta t q_j^k \tag{4.16}$$

Pictorially, we render this as a difference stencil in figure 4.3. This expresses that the function value in each point is determined by a combination of points on the previous time level.

It is convenient to summarize the set of equations (4.16) for a given k and all values of j in vector form as

$$T^{k+1} = \left(I - \frac{\alpha \Delta t}{\Delta x^2} K \right) T^k + \Delta t q^k \tag{4.17}$$

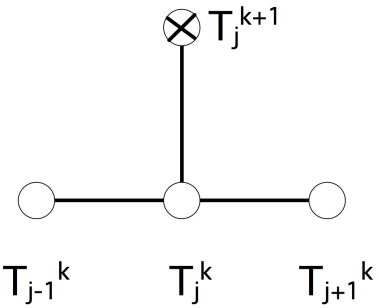


Figure 4.3: The difference stencil of the Euler forward method for the heat equation

.

where

$$K = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \end{pmatrix}$$

The important observation here is that the dominant computation for deriving the vector T^{k+1} from T^k is a simple matrix-vector multiplication:

$$T^{k+1} \leftarrow AT^k + \Delta t q^k$$

where $A = I - \frac{\alpha \Delta t}{\Delta x^2} K$. Actual computer programs using an explicit method often do not form the matrix, but evaluate the equation (4.16). However, the linear algebra formulation (4.17) is more insightful for purposes of analysis.

4.3.1.2 Implicit scheme

In equation (4.15) we let T^{k+1} be defined from T^k. We can turn this around by defining T^k from T^{k-1}, as we did for the IVP in section 4.1.2.2. For the time discretization this gives

$$\frac{\partial}{\partial t} T(x,t)\Big|_{t=t_k} \Rightarrow \frac{T_j^k - T_j^{k-1}}{\Delta t} \qquad \text{or} \qquad \frac{\partial}{\partial t} T(x,t)\Big|_{t=t_{k+1}} \Rightarrow \frac{T_j^{k+1} - T_j^k}{\Delta t}. \tag{4.18}$$

The implicit time step discretization of the whole heat equation, evaluated in t_{k+1}, now becomes:

$$\frac{T_j^{k+1} - T_j^k}{\Delta t} - \alpha \frac{T_{j-1}^{k+1} - 2T_j^{k+1} + T_{j+1}^{k+1}}{\Delta x^2} = q_j^{k+1}$$

or

$$T_j^{k+1} - \frac{\alpha \Delta t}{\Delta x^2}(T_{j-1}^k - 2T_j^k + T_{j+1}^k) = T_j^k + \Delta t q_j^k \tag{4.19}$$

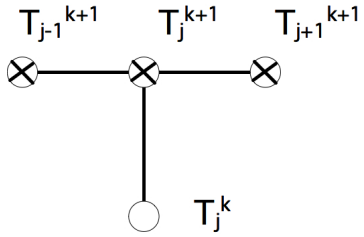


Figure 4.4: The difference stencil of the Euler backward method for the heat equation

.

Figure 4.4 renders this as a stencil; this expresses that each point on the current time level influences a combination of points on the next level. Again we express this in vector form:

$$\left(I + \frac{\alpha\Delta t}{\Delta x^2}K\right) T^{k+1} = T^k + \Delta t q^k \tag{4.20}$$

As opposed to the explicit method, where a matrix-vector multiplication sufficed, the derivation of the vector T^{k+1} from T^k now involves solving a linear system

$$T^{k+1} \leftarrow A^{-1}(T^k + \Delta t q^k)$$

where $A = I + \frac{\alpha\Delta t}{\Delta x^2}K$, a harder operation than the matrix-vector multiplication. In this case, it is not possible, as above, to evaluate the equation (4.19). Codes using an implicit method actually form the coefficient matrix, and solve the system (4.20) as such.

Exercise 4.6. Show that the flop count for a time step of the implicit method is of the same order as of the explicit method. (This only holds for a problem with one space dimension.) Give at least one argument why we consider the implicit method as computationally 'harder'.

The numerical scheme that we used here is of first order in time and second order in space: the truncation error (section 4.1.2) is $O(\Delta t + \Delta x^2)$. It would be possible to use a scheme that is second order in time by using central differences in time too. We will not consider such matters in this course.

4.3.2 Stability analysis

We now analyse the stability of the explicit and implicit schemes in a simple case. Let $q \equiv 0$, and assume $T_j^k = \beta^k e^{i\ell x_j}$ for some ℓ[5]. This assumption is intuitively defensible: since the differential equation does not 'mix' the x and t coordinates, we surmise that the solution will be a product of the separate solutions of

$$\begin{cases} u_t = c_1 u \\ u_{xx} = c_2 u \end{cases}$$

5. Actually, β is also dependent on ℓ, but we will save ourselves a bunch of subscripts, since different β values never appear together in one formula.

If the assumption holds up, we need $|\beta| < 1$ for stability.

Substituting the surmised form for T_j^k into the explicit scheme gives

$$\begin{aligned} T_j^{k+1} &= T_j^k + \frac{\alpha\Delta t}{\Delta x^2}(T_{j-1}^k - 2T_j^k + T_{j+1}^k) \\ \Rightarrow \beta^{k+1}e^{i\ell x_j} &= \beta^k e^{i\ell x_j} + \frac{\alpha\Delta t}{\Delta x^2}(\beta^k e^{i\ell x_{j-1}} - 2\beta^k e^{i\ell x_j} + \beta^k e^{i\ell x_{j+1}}) \\ &= \beta^k e^{i\ell x_j}\left[1 + \frac{\alpha\Delta t}{\Delta x^2}\left[e^{-i\ell\Delta x} - 2 + e^{i\ell\Delta x}\right]\right] \\ \Rightarrow \beta &= 1 + 2\frac{\alpha\Delta t}{\Delta x^2}[\frac{1}{2}(e^{i\ell\Delta x} + e^{-\ell\Delta x}) - 1] \\ &= 1 + 2\frac{\alpha\Delta t}{\Delta x^2}(\cos(\ell\Delta x) - 1) \end{aligned}$$

For stability we need $|\beta| < 1$:

- $\beta < 1 \Leftrightarrow 2\frac{\alpha\Delta t}{\Delta x^2}(\cos(\ell\Delta x) - 1) < 0$: this is true for any ℓ and any choice of $\Delta x, \Delta t$.
- $\beta > -1 \Leftrightarrow 2\frac{\alpha\Delta t}{\Delta x^2}(\cos(\ell\Delta x) - 1) > -2$: this is true for all ℓ only if $2\frac{\alpha\Delta t}{\Delta x^2} < 1$, that is

$$\Delta t < \frac{\Delta x^2}{2\alpha}$$

The latter condition poses a big restriction on the allowable size of the time steps: time steps have to be small enough for the method to be stable. Also, if we decide we need more accuracy in space and we halve the space discretization Δx, the number of time steps will be multiplied by four.

Let us now consider the stability of the implicit scheme. Substituting the form of the solution $T_j^k = \beta^k e^{i\ell x_j}$ into the numerical scheme gives

$$\begin{aligned} T_j^{k+1} - T_j^k &= \frac{\alpha\Delta t}{\Delta x^2}(T_{j_1}^{k+1} - 2T_j^{k+1} + T_{j+1}^{k+1}) \\ \Rightarrow \beta^{k+1}e^{i\ell\Delta x} - \beta^k e^{i\ell x_j} &= \frac{\alpha\Delta t}{\Delta x^2}(\beta^{k+1}e^{i\ell x_{j-1}} - 2\beta^{k+1}e^{i\ell x_j} + \beta^{k+1}e^{i\ell x_{j+1}}) \end{aligned}$$

Dividing out $e^{i\ell x_j}\beta^{k+1}$ gives

$$\begin{aligned} 1 &= \beta^{-1} + 2\alpha\frac{\Delta t}{\Delta x^2}(\cos\ell\Delta x - 1) \\ \beta &= \frac{1}{1 + 2\alpha\frac{\Delta t}{\Delta x^2}(1 - \cos\ell\Delta x)} \end{aligned}$$

Since $1 - \cos\ell\Delta x \in (0, 2)$, the denominator is strictly > 1. Therefore the condition $|\beta| < 1$ is always satisfied, regardless the choices of Δx and Δt: the method is always stable.

Chapter 5

Numerical linear algebra

In chapter 4 you saw how the numerical solution of partial differential equations can lead to linear algebra problems. Sometimes this is a simple problem – a matrix-vector multiplication in the case of the Euler forward method – but sometimes it is more complicated: the solution of a system of linear equations. (In other applications, which we will not discuss here, eigenvalue problems need to be solved.) You may have learned a simple algorithm for solving system of linear equations: elimination of unknowns, also called Gaussian elimination. This method can still be used, but we need some careful discussion of its efficiency. There are also other algorithms, the so-called iterative solution methods, which proceed by gradually approximating the solution of the linear system. They warrant some discussion of their own.

Because of the PDE background, we only consider linear systems that are square and nonsingular. Rectangular, in particular overdetermined, systems have important applications too in a corner of numerical analysis known as approximation theory. However, we will not cover that in this book.

5.1 Elimination of unknowns

In this section we are going to take a closer look at Gaussian elimination, or elimination of unknowns. You may have seen this method before (and if not, it will be explained below), but we will be a bit more systematic here so that we can analyze various aspects of it.

One general thread of this chapter will be the discussion of the efficiency of the various algorithms. When you learned how to solve a system of unknowns by gradually eliminating unknowns, you most likely never applied that method to a matrix larger than 4×4. The linear systems that occur in PDE solving can be thousands of times larger, and counting how many operations, as well as how much memory, they require becomes important.

The solution of a linear system can be written with a fairly simple explicit formula, using determinants. This is called '*Cramer's rule*'. It is mathematically elegant, but completely inpractical for our purposes.

If a matrix A and a vector b are given, and a vector x satisfying $Ax = b$ is wanted, then, writing $|A|$ for the determinant,

$$x_i = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1i-1} & b_1 & a_{1i+1} & \dots & a_{1n} \\ a_{21} & & \dots & & b_2 & & \dots & a_{2n} \\ \vdots & & & & \vdots & & & \vdots \\ a_{n1} & & \dots & & b_n & & \dots & a_{nn} \end{vmatrix} /|A|$$

For any matrix M the determinant is defined recursively as

$$|M| = \sum_i (-1)^i m_{1i} |M^{[1,i]}|$$

where $M^{[1,i]}$ denotes the matrix obtained by deleting row 1 and column i from M. This means that computing the determinant of a matrix of dimension n means n times computing a size $n-1$ determinant. Each of these requires $n-1$ determinants of size $n-2$, so you see that the number of operations required to compute the determinant is factorial in the matrix size. This quickly becomes prohibitive, even ignoring any issues of numerical stability. Later in this chapter you will see complexity estimates for other methods of solving systems of linear equations that are considerably more reasonable.

Let us now look at a simple of example of solving linear equations with elimination of unknowns. Consider the system

$$\begin{array}{rrrr} 6x_1 & -2x_2 & +2x_3 = & 16 \\ 12x_1 & -8x_2 & +6x_3 = & 26 \\ 3x_1 & -13x_2 & +3x_3 = & -19 \end{array}$$

We eliminate x_1 from the second and third equation by

- multiplying the first equation $\times 2$ and subtracting the result from the second equations, and
- multiplying the first equation $\times 1/2$ and subtracting the result from the third equation.

The linear system then becomes

$$\begin{array}{rrrr} 6x_1 & -2x_2 & +2x_3 = & 16 \\ 0x_1 & -4x_2 & +2x_3 = & -6 \\ 0x_1 & -12x_2 & +2x_3 = & -27 \end{array}$$

Finally, we eliminate x_2 from the third equation by multiplying the second equation by 3, and subtracting the result from the third equation:

$$\begin{array}{rrrr} 6x_1 & -2x_2 & +2x_3 = & 16 \\ 0x_1 & -4x_2 & +2x_3 = & -6 \\ 0x_1 & +0x_2 & -4x_3 = & -9 \end{array}$$

We can now solve $x_3 = 9/4$ from the last equation. Substituting that in the second equation, we get $-4x_2 = -6-2x_2 = -21/2$ so $x_2 = 21/8$. Finally, from the first equation $6x_1 = 16+2x_2-2x_3 = 16+21/4-9/2 = 76/4$ so $x_1 = 19/6$.

We can write this more compactly by omiting the x_i coefficients. Write

$$\begin{pmatrix} 6 & -2 & 2 \\ 12 & -8 & 6 \\ 3 & -13 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 16 \\ 26 \\ -19 \end{pmatrix}$$

as

$$\left[\begin{array}{ccc|c} 6 & -2 & 2 & 16 \\ 12 & -8 & 6 & 26 \\ 3 & -13 & 3 & -19 \end{array}\right] \tag{5.1}$$

then the elimination process is

$$\left[\begin{array}{ccc|c} 6 & -2 & 2 & 16 \\ 12 & -8 & 6 & 26 \\ 3 & -13 & 3 & -19 \end{array}\right] \longrightarrow \left[\begin{array}{ccc|c} 6 & -2 & 2 & 16 \\ 0 & -4 & 2 & -6 \\ 0 & -12 & 2 & -27 \end{array}\right] \longrightarrow \left[\begin{array}{ccc|c} 6 & -2 & 2 & 16 \\ 0 & -4 & 2 & -6 \\ 0 & 0 & -4 & -9 \end{array}\right]$$

In the above example, the matrix coefficients could have been any real (or, for that matter, complex) coefficients, and you could follow the elimination procedure mechanically. There is the following exception. At some point in the computation, we divided by the numbers $6, -4, -4$ which are found on the diagonal of the matrix in the last elimination step. These quantities are called the *pivots*, and clearly they are required to be nonzero. The first pivot is an element of the original matrix; the other pivots can not easily be found without doing the actual elimination.

If a pivot turns out to be zero, all is not lost for the computation: we can always exchange two matrix rows. It is not hard to show[1] that with a nonsingular matrix there is always a row exchange possible that puts a nonzero element in the pivot location.

Exercise 5.1. Suppose you want to exchange matrix rows 2 and 3 of the system of equations in equation (5.1). What other adjustments would you have to make to make sure you still compute the correct solution? What are the implications of exchanging two columns in that equation?

In general, with floating point numbers and round-off, it is very unlikely that a matrix element will become exactly zero during a computation. Also, in a PDE context, the diagonal is usually nonzero. Does that mean that pivoting is in practice almost never necessary? The answer is *no*: pivoting is desirable from a point of view of numerical stability. In the next section you will see an example that illustrates this fact.

5.2 Linear algebra in computer arithmetic

In most of this chapter, we will act as if all operations can be done in exact arithmetic. However, it is good to become aware of some of the potential problems due to our finite precision computer arithmetic. This allows us to design

1. And you can find this in any elementary linear algebra textbook.

algorithms to minimize the effect of roundoff. A more rigorous approach to the topic of numerical linear algebra includes a full-fledged error analysis of the algorithms we discuss; however, that is beyond the scope of this course; see the 'further reading' section at the end of this chapter.

Here, we will only note two paradigmatic examples of the sort of problems that can come up in computer arithmetic: we will show why 'pivoting' during LU factorization is more than a theoretical device, and we will give two examples of problems in eigenvalue calculations.

5.2.1 Roundoff control during elimination

Above, you saw that row interchanging ('pivoting') is necessary if a zero element appears on the diagonal during elimination of that row and column. Let us now look at what happens if the pivot element is not zero, but close to zero.

Consider the linear system (with ϵ presumably small)

$$\begin{pmatrix} \epsilon & 1 \\ 1 & 1 \end{pmatrix} x = \begin{pmatrix} 1+\epsilon \\ 2 \end{pmatrix}$$

which has the solution solution $x = (1,1)^t$ Using the $(1,1)$ element to clear the remainder of the first column gives:

$$\begin{pmatrix} \epsilon & 1 \\ 0 & 1-\frac{1}{\epsilon} \end{pmatrix} x = \begin{pmatrix} 1+\epsilon \\ 2-\frac{1+\epsilon}{\epsilon} \end{pmatrix}$$

We can now solve x_2 and from it x_1.

If ϵ is small, say $\epsilon < \epsilon_{\mathrm{mach}}$, the $1+\epsilon$ term in the right hand side will be simply 1: our linear system will be

$$\begin{pmatrix} \epsilon & 1 \\ 1 & 1 \end{pmatrix} x = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

but the solution will still be $(1,1)^t$.

Next, $1/\epsilon$ will be very large, so the second component of the right hand side after elimination will be $2-\frac{1}{\epsilon} = -1/\epsilon$. Also, the $(2,2)$ element of the matrix is then $-1/\epsilon$ instead of $1-1/\epsilon$:

$$\begin{pmatrix} \epsilon & 1 \\ 0 & -1/\epsilon \end{pmatrix} x = \begin{pmatrix} 1 \\ -1/\epsilon \end{pmatrix}$$

We get the correct value $x_2 = 1$, but

$$\epsilon x_1 + x_2 = 1 \Rightarrow \epsilon x_1 = 0 \Rightarrow x_1 = 0,$$

which is 100% wrong, or infinitely wrong depending on how you look at it.

What would have happened if we had pivoted as described above? We exchange the matrix rows, giving

$$\begin{pmatrix} 1 & 1 \\ \epsilon & 1 \end{pmatrix} x = \begin{pmatrix} 2 \\ 1+\epsilon \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 1 \\ 0 & 1-\epsilon \end{pmatrix} x = \begin{pmatrix} 2 \\ 1-\epsilon \end{pmatrix}$$

Now we get, regardless the size of epsilon:

$$x_1 = \frac{1-\epsilon}{1-\epsilon} = 1, \qquad x_2 = 2 - x_1 = 1$$

In this example we used a very small value of ϵ; a much more refined analysis shows that even with ϵ greater than the machine precision pivoting still makes sense. The general rule of thumb is: *Always do row exchanges to get the largest remaining element in the current column into the pivot position.* In chapter 4 you will see matrices that arise in certain practical applications; it can be shown that for them pivoting is never necessary; see exercise **??**.

The pivoting that was discussed above is also known as *partial pivoting*, since it is based on row exchanges only. Another option would be *full pivoting*, where row and column exchanges are combined to find the largest element in the remaining subblock, to be used as pivot. Finally, *diagonal pivoting* applies the same exchange to rows and columns. This means that pivots are only searched on the diagonal. From now on we will only consider partial pivoting.

5.2.2 Influence of roundoff on eigenvalue computations

Consider the matrix

$$A = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}$$

where $\epsilon_{\mathrm{mach}} < |\epsilon| < \sqrt{\epsilon_{\mathrm{mach}}}$, which has eigenvalues $1+\epsilon$ and $1-\epsilon$. If we calculate its characteristic polynomial in computer arithmetic

$$\begin{vmatrix} 1-\lambda & \epsilon \\ \epsilon & 1-\lambda \end{vmatrix} = \lambda^2 - 2\lambda + (1-\epsilon^2)\lambda^2 - 2\lambda + 1.$$

we find a double eigenvalue 1. Note that the exact eigenvalues are expressible in working precision; it is the algorithm that causes the error. Clearly, using the characteristic polynomial is not the right way to compute eigenvalues, even in well-behaved, symmetric positive definite, matrices.

An unsymmetric example: let A be the matrix of size 20

$$A = \begin{pmatrix} 20 & 20 & & & \emptyset \\ & 19 & 20 & & \\ & & \ddots & \ddots & \\ & & & 2 & 20 \\ \emptyset & & & & 1 \end{pmatrix}.$$

Since this is a triangular matrix, its eigenvalues are the diagonal elements. If we perturb this matrix by setting $A_{20,1} = 10^{-6}$ we find a perturbation in the eigenvalues that is much larger that in the elements:

$$\lambda = 20.6 \pm 1.9i, 20.0 \pm 3.8i, 21.2, 16.6 \pm 5.4i, \ldots$$

5.3 LU factorization

So far, we have looked at eliminating unknowns in the context of solving a single system of linear equations. Suppose you need to solve more than one system with the same matrix, but with different right hand sides. Can you use any of the work you did in the first system to make solving subsequent ones easier?

The answer is yes. You can split the solution process in a part that only concerns the matrix, and a part that is specific to the right hand side. If you have a series of systems to solve, you have to do the first part only once, and, luckily, that even turns out to be the larger part of the work.

Let us take a look at the same example of section 5.1 again.

$$A = \begin{pmatrix} 6 & -2 & 2 \\ 12 & -8 & 6 \\ 3 & -13 & 3 \end{pmatrix}$$

In the elimination process, we took the 2nd row minus $2\times$ the first and the 3rd row minus $1/2\times$ the first. Convince yourself that this combining of rows can be done by multiplying A from the left by

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix}$$

which is the identity with the elimination coefficients in the first column, below the diagonal. You see that the first step in elimination of variables is equivalent to transforming the system $Ax = b$ to $L_1 Ax = L_1 b$.

In the next step, you subtracted $3\times$ the second row from the third. Convince yourself that this corresponds to left-multiplying the current matrix $L_1 A$ by

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3 & 1 \end{pmatrix}$$

We have now transformed our system $Ax = b$ into $L_2 L_1 Ax = L_2 L_1 b$, and $L_2 L_1 A$ is of 'upper triangular' form. If we define $U = L_2 L_1 A$, then $A = L_1^{-1} L_2^{-1} U$. How hard is it to compute matrices such as L_2^{-1}? Remarkably easy, it turns out to be.

We make the following observations:

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix} \qquad L_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}$$

and likewise

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3 & 1 \end{pmatrix} \qquad L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix}$$

and even more remarkable:

$$L_1^{-1}L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1/2 & 3 & 1 \end{pmatrix},$$

that is, $L_1^{-1}L_2^{-1}$ contains the elements of L_1^{-1}, L_2^{-1} unchanged.

If we define $L = L_1^{-1}L_2^{-1}$, we now have $A = LU$; this is called an *LU factorization*. We see that the coefficients of L below the diagonal are the negative of the coefficients used during elimination. Even better, the first column of L can be written while the first column of A is being eliminated, so the computation of L and U can be done without extra storage, at least if we can afford to lose A.

5.3.1 The algorithm

Let us write out the LU factorization algorithm in more or less formal code.

```
⟨LU factorization⟩:
    for k = 1, n − 1:
    ⟨eliminate values in column k⟩
⟨eliminate values in column k⟩:
    for i = k + 1 to n:
      ⟨compute multiplier for row i⟩
      ⟨update row i⟩
⟨compute multiplier for row i⟩
    a_ik ← a_ik/a_kk
⟨update row i⟩:
    for j = k + 1 to n:
      a_ij ← a_ij − a_ik ∗ a_kj
```

Or, putting everything together:

```
⟨LU factorization⟩:
    for k = 1, n − 1:
      for i = k + 1 to n:
        a_ik ← a_ik/a_kk
        for j = k + 1 to n:
          a_ij ← a_ij − a_ik ∗ a_kj
```
(5.2)

This is the most common way of presenting the LU factorization. However, other ways of computing the same result exist; see section 5.6.

5.3.2 Uniqueness

It is always a good idea, when discussing numerical algorithms, to wonder if different ways of computing lead to the same result. This is referred to as the 'uniqueness' of the result, and it is of practical use: if the computed result is unique, swapping one software library for another will not change anything in the computation.

Let us consider the uniqueness of LU factorization. The definition of an LU factorization algorithm (without pivoting) is that, given a nonsingular matrix A, it will give a lower triangular matrix L and upper triangular matrix U such that $A = LU$. The above algorithm for computing an LU factorization is deterministic (it does not contain instructions 'take any row that satisfies...'), so given the same input, it will always compute the same output. However, other algorithms are possible, so we need worry whether they give the same result.

Let us then assume that $A = L_1U_1 = L_2U_2$ where L_1, L_2 are lower triangular and U_1, U_2 are upper triangular. Then, $L_2^{-1}L_1 = U_2U_1^{-1}$. In that equation, the left hand side is the product of lower triangular matrices, while the right hand side contains only upper triangular matrices.

Exercise 5.2. Prove that the product of lower triangular matrices is lower triangular, and the product of upper triangular matrices upper triangular. Is a similar statement true for inverses of nonsingular triangular matrices?

The product $L_2^{-1}L_1$ is apparently both lower triangular and upper triangular, so it must be diagonal. Let us call it D, then $L_1 = L_2D$ and $U_2 = DU_1$. The conclusion is that LU factorization is not unique, but it *is* unique 'up to diagonal scaling'.

Exercise 5.3. The algorithm in section 5.3.1 resulted in a lower triangular factor L that had ones on the diagonal. Show that this extra condition make the factorization unique.

Exercise 5.4. Show that an alternative condition of having ones on the diagonal of U is also sufficient for the uniqueness of the factorization.

Since we can demand a unit diagonal in L or in U, you may wonder if it is possible to have both. (Give a simple argument why this is not strictly possible.) We can do the following: suppose that $A = LU$ where L and U are nonsingular lower and upper triangular, but not normalized in any way. Write

$$L = (I + L')D_L, \qquad U = D_U(I + U'), \qquad D = D_LD_U.$$

After some renaming we now have a factorization

$$A = (I + L)D(I + U) \tag{5.3}$$

Exercise 5.5. Show that you can also normalize the factorization on the form

$$A = (D + L)D^{-1}(D + U).$$

Consider the factorization of a tridiagonal matrix this way. How do L and U relate to the triangular parts of A?

5.3.3 Pivoting

Above, you saw examples where pivoting, that is, exchanging rows, was necessary during the factorization process, either to guarantee the existence of a nonzero pivot, or for numerical stability. We will now integrate pivoting into the LU factorization.

Let us first observe that row exchanges can be described by a matrix multiplication. Let

$$P^{(i,j)} = \begin{array}{c} \\ i \\ \\ \\ j \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & & i & & j & \end{array} \\ \begin{pmatrix} 1 & 0 & & & & & \\ 0 & \ddots & \ddots & & & & \\ & & 0 & & 1 & & \\ & & & I & & & \\ & & 1 & & 0 & & \\ & & & & & I & \\ & & & & & & \ddots \end{pmatrix} \end{array}$$

then $P^{(i,j)}A$ is the matrix A with rows i and j exchanged. Since we may have to pivot in every iteration of the factorization process, we introduce a sequence p_i containing the j values, and we write $P^{(i)} \equiv P^{(i,p(i))}$ for short.

Exercise 5.6. Show that $P^{(i)}$ is its own inverse.

The process of factoring with partial pivoting can now be described as:

- Let $A^{(i)}$ be the matrix with columns $1 \ldots i-1$ eliminated, and partial pivoting applied to get the right element in the (i,i) location.
- Let $\ell^{(i)}$ be the vector of multipliers in the i-th elimination step. (That is, the elimination matrix L_i in this step is the identity plus $\ell^{(i)}$ in the i-th column.)
- Let $P^{(i+1)}$ (with $j \geq i+1$) be the matrix that does the partial pivoting for the next elimination step as described above.
- Then $A^{(i+1)} = P^{(i+1)}L_iA^{(i)}$.

In this way we get a factorization of the form

$$L_{n-1}P^{(n-2)}L_{n-2}\cdots L_1P^{(1)}A = U.$$

Suddenly it has become impossible to write $A = LU$: we would have to write

$$A = P^{(1)}L_1^{-1}\cdots P^{(n-2)}L_{n-1}^{-1}U. \tag{5.4}$$

Exercise 5.7. Recall from sections 1.5.7 and 1.5.8 that blocked algorithms are often desirable from a performance point of view. Why is the 'LU factorization with interleaved pivoting matrices' in equation (5.4) bad news for performance?

Fortunately, equation (5.4) can be simplified: the P and L matrices 'almost commute'. We show this by looking at an example: $P^{(2)}L_1 = \tilde{L}_1 P^{(2)}$ where $\tilde{L}_1$ is very close to L_1.

$$\begin{pmatrix} 1 & & & \\ & 0 & & 1 \\ & & I & \\ & 1 & & 0 \\ & & & & I \end{pmatrix} \begin{pmatrix} 1 & & \emptyset & \\ \vdots & 1 & & \\ \ell^{(1)} & & \ddots & \\ \vdots & & & 1 \end{pmatrix} = \begin{pmatrix} 1 & & \emptyset & \\ \vdots & 0 & & 1 \\ \tilde{\ell}^{(1)} & & & \\ \vdots & 1 & & 0 \\ \vdots & & & & I \end{pmatrix} = \begin{pmatrix} 1 & & \emptyset & \\ \vdots & 1 & & \\ \tilde{\ell}^{(1)} & & \ddots & \\ \vdots & & & 1 \end{pmatrix} \begin{pmatrix} 1 & & & \\ & 0 & & 1 \\ & & I & \\ & 1 & & 0 \\ & & & & I \end{pmatrix}$$

where $\tilde{\ell}^{(1)}$ is the same as $\ell^{(1)}$, except that elements i and $p(i)$ have been swapped. You can now easily convince yourself that similarly $P^{(2)}$ et cetera can be 'pulled through' L_1.

As a result we get

$$P^{(n-2)} \cdots P^{(1)} A = \tilde{L}_1^{-1} \cdots L_{n-1}^{-1} U = \tilde{L} U. \tag{5.5}$$

This means that we can again form a matrix L just as before, except that every time we pivot, we need to update the columns of L that have already been computed.

Exercise 5.8. If we write equation (5.5) as $PA = LU$, we get $A = P^{-1}LU$. Can you come up with a simple formula for P^{-1} in terms of just P? Hint: each $P^{(i)}$ is symmetric and its own inverse; see the exercise above.

Exercise 5.9. Earlier, you saw that 2D BVP (section 4.2.2.2) give rise to a certain kind of matrix. We stated, without proof, that for these matrices pivoting is not needed. We can now formally prove this, focusing on the crucial property of *diagonal dominance*:

$$\forall_i a_{ii} \geq \sum_{j \neq i} |a_{ij}|.$$

Assume that a matrix A satisfies $\forall_{j \neq i} \colon a_{ij} \leq 0$. Show that the matrix is diagonally dominant iff there are vectors $u, v \geq 0$ (meaning that each component is nonnegative) such that $Au = v$. Show that, after eliminating a variable, for the remaining matrix $\tilde{A}$ there are again vectors $\tilde{u}, \tilde{v} \geq 0$ such that $\tilde{A}\tilde{u} = \tilde{v}$.

Now finish the argument that (partial) pivoting is not necessary if A is symmetric and diagonally dominant. (One can actually prove that pivoting is not necessary for any symmetric positive definite (SPD) matrix, and diagonal dominance is a stronger condition than SPD-ness.)

5.3.4 Solving the system

Now that we have a factorization $A = LU$, we can use this to solve the linear system $Ax = LUx = b$. If we introduce a temporary vector $y = Ux$, then we see this takes two steps:

$$Ly = b, \qquad Ux = z.$$

The first part, $Ly = b$ is called the 'lower triangular solve', since it involves the lower triangular matrix L.

$$\begin{pmatrix} 1 & & & & \emptyset \\ \ell_{21} & 1 & & & \\ \ell_{31} & \ell_{32} & 1 & & \\ \vdots & & & \ddots & \\ \ell_{n1} & \ell_{n2} & & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \\ \vdots \\ b_n \end{pmatrix}$$

In the first row, you see that $y_1 = b_1$. Then, in the second row $\ell_{21}y_1 + y_2 = b_2$, so $y_2 = b_2 - \ell_{21}y_1$. You can imagine how this continues: in every i-th row you can compute y_i from the previous y-values:

$$y_i = b_i - \sum_{j<i} \ell_{ij}y_j.$$

Since we compute y_i in increasing order, this is also known as the 'forward sweep'.

The second half of the solution process, the 'upper triangular solve', or 'backward sweep' computes x from $Ux = y$:

$$\begin{pmatrix} u_{11} & u_{12} & \ldots & u_{1n} \\ & u_{22} & \ldots & u_{2n} \\ & & \ddots & \vdots \\ \emptyset & & & u_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Now we look at the last line, which immediately tells $x_n = u_{nn}^{-1}y_n$. From this, the line before the last states $u_{n-1n-1}x_{n-1} + u_{n-1n}x_n = y_{n-1}$, which gives $x_{n-1} = u_{n-1n-1}^{-1}(y_{n-1} - u_{n-1n}x_n)$. In general, we can compute

$$x_i = u_{ii}^{-1}(y_i - \sum_{j>i} u_{ij}y_j)$$

for decreasing values of i.

Exercise 5.10. In the backward sweep you had to divide by the numbers u_{ii}. That is not possible if any of them are zero. Relate this problem back to the above discussion. What reason was given there to explain that this will never be a problem?

5.3.5 Complexity

In the beginning of this chapter, we indicated that not every method for solving a linear system takes the same number of operations. Let us therefore take a closer look at the complexity[2], that is, the number of operations as function of the problem size, of the use of an LU factorization in solving the linear system.

The complexity of solving the linear system, given the LU factorization, is easy to compute. Looking at the lower and upper triangular part together, you see that you perform a multiplication with all off-diagonal elements (that is,

2. See appendix A.2 for an explanation of complexity.

elements ℓ_{ij} or u_{ij} with $i \neq j$). Furthermore, the upper triangular solve involves divisions by the u_{ii} elements. Now, division operations are in general much more expensive than multiplications, so in this case you would compute the values $1/u_{ii}$, and store them instead.

Summing up, you see that, on a system of size $n \times n$, you perform n^2 multiplications and roughly the same number of additions. This is the same complexity as of a simple matrix-vector multiplication, that is, of computing Ax given A and x.

The complexity of computing the LU factorization is a bit more involved to compute. Refer to the algorithm in section 5.3.1. You see that in the k-th step two things happen: the computation of the multipliers, and the updating of the rows. There are $n - k$ multipliers to be computed, each of which involve a division. After that, the update takes $(n-k)^2$ additions and multiplications. If we ignore the divisions for now, because there are fewer of them, we find that the LU factorization takes $\sum_{k=1}^{n-1} 2(n-k)^2$ operations. If we number the terms in this sum in the reverse order, we find

$$\#\text{ops} = \sum_{k=1}^{n-1} 2k^2.$$

Since we can approximate a sum by an integral, we find that this is $2/3n^3$ plus some lower order terms. This is of a higher order than solving the linear system: as the system size grows, the cost of constructing the LU factorization completely dominates.

5.3.6 Accuracy

In section 5.2 you saw some simple examples of the problems that stem from the use of computer arithmetic, and how these motivated the use of pivoting. Even with pivoting, however, we still need to worry about the accumulated effect of roundoff errors. A productive way of looking at the question of attainable accuracy is to consider that by solving a system $Ax = b$ we get a numerical solution $x + \Delta x$ which is the exact solution of a slightly different linear system:

$$(A + \Delta A)(x + \Delta x) = b + \Delta b.$$

Analyzing these statements quickly leads to bounds such as

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{2\epsilon\kappa(A)}{1 - \epsilon\kappa(A)}$$

where ϵ is the *machine precision* (see section 3.2.3) and $\kappa(A) = \|A\|\|A^{-1}\|$ is called the *condition number* of the matrix A (see appendix A.1). Without going into this in any detail, we remark that the condition number is related to eigenvalues (strictly speaking: singular values) of the matrix.

The analysis of the accuracy of algorithms is a field of study in itself; see for instance the book by Higham [51].

5.3.7 Block algorithms

Often, matrices have a natural block structure, such as in the case of two-dimensional BVPs; section 4.2.2.2. Many linear algebra operations can be formulated in terms of these blocks. For this we write a matrix as

$$A = \begin{pmatrix} A_{11} & \ldots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \ldots & A_{MN} \end{pmatrix}$$

where M, N are the block dimensions, that is, the dimension expressed in terms of the subblocks. Usually, we choose the blocks such that $M = N$ and the diagonal blocks are square.

As a simple example, consider the matrix-vector product $y = Ax$, expressed in block terms.

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_M \end{pmatrix} = \begin{pmatrix} A_{11} & \ldots & A_{1M} \\ \vdots & & \vdots \\ A_{M1} & \ldots & A_{MM} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_M \end{pmatrix}$$

To see that the block algorithm computes the same result as the old scalar algorithm, we look at a component X_{i_k}, that is the k-th scalar component of the i-th block. First,

$$Y_i = \sum_j A_{ij} X_j$$

so

$$Y_{i_k} = \big(\sum_j A_{ij} X_j\big)_k = \sum_j (A_{ij} X_j)_k = \sum_j \sum_\ell A_{ij_{k\ell}} X_{j_\ell}$$

which is the product of the k-th row of the i-th blockrow of A with the whole of X.

A more interesting algorithm is the block version of the LU factorization. The algorithm (5.2) then becomes

$$\begin{array}{l} \langle LU \text{ factorization} \rangle: \\ \quad \text{for } k = 1, n-1: \\ \quad\quad \text{for } i = k+1 \text{ to } n: \\ \quad\quad\quad A_{ik} \leftarrow A_{ik} A_{kk}^{-1} \\ \quad\quad\quad \text{for } j = k+1 \text{ to } n: \\ \quad\quad\quad\quad A_{ij} \leftarrow A_{ij} - A_{ik} \cdot A_{kj} \end{array} \tag{5.6}$$

which mostly differs from the earlier algorithm in that the division by a_{kk} has been replaced by a multiplication by A_{kk}^{-1}. Also, the U factor will now have pivot blocks, rather than pivot elements, on the diagonal, so U is only 'block upper triangular', and not strictly upper triangular.

Exercise 5.11. We would like to show that the block algorithm here again computes the same result as the scalar one. Doing so by looking explicitly at the computed elements is cumbersome, so we take another approach. First, recall from section 5.3.2 that LU factorizations are unique: if $A = L_1U_1 = L_2U_2$ and L_1, L_2 have unit diagonal, then $L_1 = L_2$, $U_1 = U_2$.
Next, consider the computation of A_{kk}^{-1}. Show that this can be done easily by first computing an LU factorization of A_{kk}. Now use this to show that the block LU factorization can give L and U factors that are strictly triangular. The uniqueness of LU factorizations then proves that the block algorithm computes the scalar result.

Block algorithms are interesting for a variety of reasons. On single processors they are the key to high cache utilization; see section 1.5.6. On shared memory architectures, including current *multicore* processors, they can be used to schedule parallel tasks on the processors/cores; see section 6.10.

5.4 Sparse matrices

In section 4.2.2.2 you saw that the discretization of BVPs (and IBVPs) may give rise to sparse matrices. Since such a matrix has n^2 elements but only $O(n)$ nonzeros, it would be a big waste of space to store this as a square array. Additionally, we want to avoid operating on zero elements.

In this section we will explore what form familiar linear algebra operations take when applied to sparse matrices. First we will concern ourselves with describing some of the ways to store a sparse matrix.

5.4.1 Storage of sparse matrices

It is pointless to look for an exact definition of *sparse matrix*, but an operational definition is that a matrix is called 'sparse' if there are enough zeros to make specialized storage feasible. We will discuss here briefly the most popular storage schemes for sparse matrices. Since a matrix is no longer stored as a simple 2-dimensional array, algorithms using such storage schemes need to be rewritten too.

5.4.1.1 Diagonal storage

In section 4.2.2.1 you have seen examples of sparse matrices that were banded. In fact, their nonzero elements are located precisely on a number of subdiagonals. For such a matrix, a specialized storage scheme is possible.

Let us take as an example the matrix of the one-dimensional BVP (section 4.2.2.1). Its elements are located on three subdiagonals: the main diagonal and the first super and subdiagonal. The idea of *storage by diagonals* or *diagonal storage* is to store the diagonals consecutively in memory. The most economical storage scheme for such a matrix would store the $2n - 2$ elements consecutively. However, for various reasons it is more convenient to waste a few storage locations, as shown in figure 5.1.

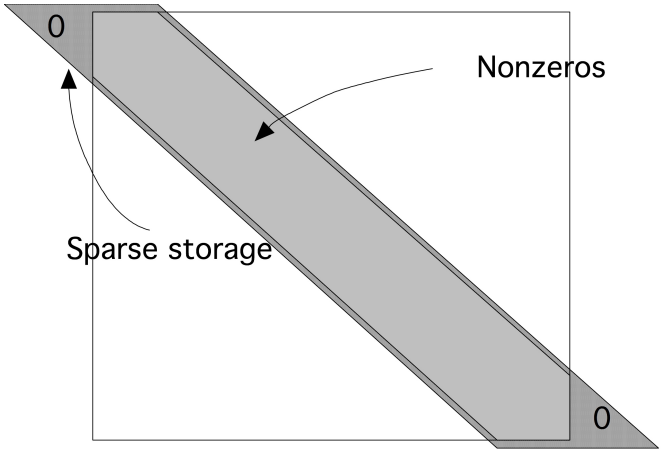


Figure 5.1: Diagonal storage of a banded matrix

Thus, for a matrix with size $n \times n$ and a *bandwidth* p, we need a rectangular array of size $n \times p$ to store the matrix. The matrix of equation (4.9) would then be stored as

$$\begin{bmatrix} \star & 2 & -1 \\ -1 & 2 & -1 \\ \vdots & \vdots & \vdots \\ -1 & 2 & \star \end{bmatrix} \tag{5.7}$$

where the stars indicate array elements that do not correspond to matrix elements: they are the triangles in the top left and bottom right in figure 5.1.

If we apply this scheme to the matrix of the two-dimensional BVP (section 4.2.2.2), it becomes wasteful, since we would be storing many zeros that exist inside the band. Therefore, we refine this scheme by storing only the nonzero diagonals: if the matrix has p nonzero diagonals, we need an $n \times p$ array. For the matrix of equation (4.14) this means:

$$\begin{bmatrix} \star & \star & 4 & -1 & -1 \\ \vdots & \vdots & 4 & -1 & -1 \\ \vdots & -1 & 4 & -1 & -1 \\ -1 & -1 & 4 & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & 4 & \star & \star \end{bmatrix}$$

Of course, we need an additional integer array telling us the locations of these nonzero diagonals.

5.4.1.2 Operations on diagonal storage

The most important operation on sparse matrices is the matrix-vector product. With a matrix stored by diagonals, as described above, it is still possible to perform the ordinary rowwise product. However, with a small bandwidth, this gives short vector lengths and relatively high loop overhead, so it will not be efficient. It is possible do to much better than that.

If we look at how the matrix elements are used in the matrix-vector product, then we see that the main diagonal is used as

$$y_i \leftarrow y_i + A_{ii}x_i,$$

the first superdiagonal is used as

$$y_i \leftarrow y_i + A_{ii+1}x_{i+1} \quad \text{for } i < n,$$

and the first subdiagonal as

$$y_i \leftarrow y_i + A_{ii-1}x_{i-1} \quad \text{for } i > 1.$$

In other words, the whole matrix-vector product can be executed in just three vector operations of length n (or $n-1$), instead of n inner products of length 3 (or 2).

```
for diag = -diag_left, diag_right
    for loc = max(1,1-diag), min(n,n-diag)
        y(loc) = y(loc) + val(loc,diag) * x(loc+diag)
    end
end
```

Exercise 5.12. Write a routine that computes $y \leftarrow A^t x$ by diagonals. Implement it in your favourite language and test it on a random matrix.

5.4.1.3 Compressed row storage

If we have a sparse matrix that does not have a simple band structure, or where the number of nonzero diagonals becomes impractically large, we use the more general Compressed Row Storage (CRS) scheme. Consider an example of a sparse matrix:

$$A = \begin{pmatrix} 10 & 0 & 0 & 0 & -2 & 0 \\ 3 & 9 & 0 & 0 & 0 & 3 \\ 0 & 7 & 8 & 7 & 0 & 0 \\ 3 & 0 & 8 & 7 & 5 & 0 \\ 0 & 8 & 0 & 9 & 9 & 13 \\ 0 & 4 & 0 & 0 & 2 & -1 \end{pmatrix}. \tag{5.8}$$

In the so-called *CRS* we store all nonzeros by row in a single real array. The column indices are similarly stored in an integer array, and we store pointers to where the columns start. Using 0-based indexing this gives:

val	10	-2	3	9	3	7	8	7	$3 \cdots 9$	13	4	2	-1
col_ind	0	4	0	1	5	1	2	3	$0 \cdots 4$	5	1	4	5

row_ptr	0	2	5	8	12	16	19

.

A simple variant of CRS is *Compressed Column Storage (CCS)* where the elements in columns are stored contiguously. Another storage scheme you may come across is *coordinate storage*, where the matrix is stored as a list of triplets $\langle i, j, a_{ij} \rangle$.

Finally, the ultimately flexible storage scheme is Coordinate Storage (COO) that stores a list of (i, j, a_{ij}) triples. The popular Matrix Market website [66] uses a variant of this scheme.

5.4.1.4 Algorithms on compressed row storage

In this section we will look at the form some algorithms take in CRS.

The most common operation is the matrix-vector product:

```
for (row=0; row<nrows; row++) {
   s = 0;
   for (icol=ptr[row]; icol<ptr[row+1]; icol++) {
      int col = ind[icol];
      s += a[icol] * x[col];
   }
   y[row] = s;
}
```

You recognize the standard matrix-vector product algorithm for $y = Ax$, where the inner product is taken of each row A_{i*} and the input vector x.

Now, how about if you wanted to compute the product $y = A^t x$? In that case you need rows of A^t, or, equivalently, columns of A. Finding arbitrary columns of A is hard, requiring lots of searching, so you may think that this algorithm is correspondingly hard to compute. Fortunately, that is not true.

If we exchange the i and j loop in the standard algorithm for $y = Ax$, we get

$$\begin{array}{l} y \leftarrow 0 \\ \text{for } i: \\ \quad \text{for } j: \\ \quad\quad y_i \leftarrow y_i + a_{ij} x_j \end{array} \quad \Rightarrow \quad \begin{array}{l} y \leftarrow 0 \\ \text{for } j: \\ \quad \text{for } i: \\ \quad\quad y_i \leftarrow y_i + a_{ij} x_j \end{array}$$

We see that in the second variant, columns of A are accessed, rather than rows. This means that we can use the second algorithm for computing the $A^t x$ product by rows.

Exercise 5.13. Write out the code for the transpose product $y = A^t x$ where A is stored in CRS format. Write a simple test program and confirm that your code computes the right thing.

Exercise 5.14. What if you need access to both rows and columns at the same time? Implement an algorithm that tests whether a matrix stored in CRS format is symmetric. Hint: keep an array of pointers, one for each row, that keeps track of how far you have progressed in that row.

Exercise 5.15. The operations described so far are fairly simple, in that they never make changes to the sparsity structure of the matrix. The CRS format, as described above, does not allow you to add new nonzeros to the matrix, but it is not hard to make an extension that does allow it.
Let numbers $p_i, i = 1 \ldots n$, describing the number of nonzeros in the i-th row, be given. Design an extension to CRS that gives each row space for q extra elements. Implement this scheme and test it: construct a matrix with p_i nonzeros in the i-th row, and check the correctness of the matrix-vector product before and after adding new elements, up to q elements per row.
Now assume that the matrix will never have more than a total of qn nonzeros. Alter your code so that it can deal with starting with an empty matrix, and gradually adding nonzeros in random places. Again, check the correctness.

5.4.2 Sparse matrices and graph theory

Many arguments regarding sparse matrices can be formulated in terms of graph theory. (If you are not familiar with the basics of graph theory, see appendix A.6.) To see why this can be done, consider a sparse matrix A of size n and observe that we can define a graph $\langle E, V\rangle$ by $V = \{1, \ldots, n\}$, $E = \{(i, j) \colon a_{ij} \neq 0\}$. For simplicity, we assume that A has a nonzero diagonal. If necessary, we can attach weights to this graph, defined by $w_{ij} = a_{ij}$. The graph is then denoted $\langle E, V, W\rangle$.

Graph properties now correspond to matrix properties; for instance, the degree of the graph is the maximum number of nonzeros per row, not counting the diagonal element. As another example, if the graph of the matrix is an *undirected graph*, this means that $a_{ij} \neq 0 \Leftrightarrow a_{ji} \neq 0$. We call such a matrix *structurally symmetric*: it is not truly symmetric in the sense that $\forall_{ij} \colon a_{ij} = a_{ij}$, but every nonzero in the upper triangle is mirrored in the lower triangle and vice versa.

One advantage of considering the graph of a matrix is that graph properties do not depend on how we order the nodes, that is, they are invariant under permutation of the matrix.

Exercise 5.16. Let us take a look at what happens with a matrix A when the nodes of its graph $G = \langle V, E, W\rangle$ are renumbered. As a simple example, we number the nodes backwards; that is, with n the number of nodes, we map node i to $n+1-i$. Correspondingly, we find a new graph $G' = \langle V, E', W'\rangle$ where

$$(i, j) \in E' \Leftrightarrow (n + 1 - i, n + 1 - j) \in E, \qquad w'_{ij} = w_{n+1-i,n+1-j}.$$

What does this renumbering imply for the matrix A' that corresponds to G'? If you exchange the labels i, j on two nodes, what is the effect on the matrix A?

Some graph properties can be hard to see from the sparsity pattern of a matrix, but are easier deduced from the graph.

Exercise 5.17. Let A be a tridiagonal matrix (see section 4.2.2.1) of size n with n odd. What does the graph of A look like? Now zero the offdiagonal elements closest to the 'middle' of the matrix: let $a_{(n+1)/2,(n+1)/2+1} = a_{(n+1)/2+1,(n+1)/2} = 0$. Describe what that does to the graph of A. Such a graph is called *reducible*. Consider the permutation that results from putting the nodes in the following sequence: $1, 3, 5, \ldots, (n+1)/2, 2, 4, 6, \ldots, (n-1)/2$. What does the sparsity pattern of the permuted matrix look like? Note that the reducibility of the graph is now harder to read from the sparsity pattern.

5.4.3 LU factorizations of sparse matrices

In section 4.2.2.1 the one-dimensional BVP led to a linear system with a tridiagonal coefficient matrix. If we do one step of Gaussian elimination, the only element that needs to be eliminated is in the second row:

$$\begin{pmatrix} 2 & -1 & 0 & \ldots & \\ -1 & 2 & -1 & & \\ 0 & -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \Rightarrow \left(\begin{array}{c|cccc} 2 & -1 & 0 & \ldots & \\ \hline 0 & 2-\frac{1}{2} & -1 & & \\ 0 & -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots & \ddots \end{array}\right)$$

There are two important observations to be made: one is that this elimination step does not change any zero elements to nonzero. The other observation is that the part of the matrix that is left to be eliminated is again tridiagonal. Inductively, during the elimination no zero elements change to nonzero: the sparsity pattern of $L + U$ is the same as of A, and so the factorization takes the same amount of space to store as the matrix.

The case of tridiagonal matrices is unfortunately not typical, as we will now see in the case of two-dimensional problems. We write such matrices of size $N \times N$ as block matrices with block dimension n, each block being of size n. (Refresher question: where do these blocks come from?) Now, in the first elimination step we need to zero two elements, a_{21} and $a_{n+1,1}$.

$$\left(\begin{array}{ccccc:cc:c} 4 & -1 & 0 & \ldots & & -1 & & \\ -1 & 4 & -1 & 0 & \ldots & 0 & -1 & \\ & \ddots & \ddots & \ddots & & & \ddots & \\ \hdashline -1 & 0 & \ldots & & & 4 & -1 & \\ 0 & -1 & 0 & \ldots & & -1 & 4 & -1 \end{array}\right) \Rightarrow \left(\begin{array}{c|cccc:c:c} 4 & -1 & 0 & \ldots & & -1 & \\ \hline & 4-\frac{1}{4} & -1 & 0 & \ldots & -1/4 & -1 & \\ & \ddots & \ddots & \ddots & & & \ddots & \\ \hdashline & -1/4 & \ldots & & & 4-\frac{1}{4} & -1 & \\ & -1 & 0 & \ldots & & -1 & 4 & -1 \end{array}\right)$$

You see that eliminating $a_{n+1,1}$ causes two *fill* elements to appear: $a_{2,n+1} = 0$, but in the modified matrix $\tilde{a}_{2,n+1} \neq 0$. We define *fill locations* as locations (i, j) where $a_{ij} = 0$, but $(L + U)_{ij} \neq 0$.

Exercise 5.18. How many fill elements are there in the next eliminating step? Can you characterize for the whole of the factorization which locations in A get filled in $L + U$?

Exercise 5.19. The LAPACK software for dense linear algebra has an LU factorization routine that overwrites the input matrix with the factors. Above you saw that is possible since the columns of L are generated precisely as the columns of A are eliminated. Why is such an algorithm not possible if the matrix is stored in sparse format?

5.4.3.1 Fill-in estimates

In the above example you saw that the factorization of a sparse matrix can take much more space than the matrix itself, but still less than storing an entire array of size the matrix dimension. We will now give some bounds for the *space complexity* of the factorization, that is, amount of space needed to execute the factorization algorihm.

Exercise 5.20. Prove the following statements.

1. Assume that the matrix A has a *halfbandwidth* p, that is, $a_{ij} = 0$ if $|i - j| > p$. Show that, after a factorization without pivoting, $L + U$ has the same halfbandwidth.
2. Show that, after a factorization with partial pivoting, L has a *left halfbandwidth* of p, that is $\ell_{ij} = 0$ if $i > j + p$, whereas U has a *right halfbandwidth* of $2p$, that is, $u_{ij} = 0$ if $j > i + 2p$.
3. Assuming no pivoting, show that the fill-in can be characterized as follows:

   Consider row i. Let $j_{\min}$ be the leftmost nonzero in row i, that is $a_{ij} = 0$ for $j < j_{\min}$. Then there will be no fill-in in row i to the left of column $j_{\min}$. Likewise, if $i_{\min}$ is the topmost nonzero in column j, there will be no fill-in in column j above row $i_{\min}$.

   As a result, L and U have a 'skyline' profile. Given a sparse matrix, it is now easy to allocate enough storage to fit a factorization without pivoting.

This exercise shows that we can allocate enough storage for the factorization of a banded matrix:

- for the factorization without pivoting of a matrix with bandwidth p, an array of size $N \times p$ suffices;
- the factorization with partial pivoting of a matrix left halfbandwidth p and right halfbandwidth q can be stored in $N \times (p + 2q + 1)$.
- A skyline profile, sufficient for storing the factorization, can be constructed based on the specific matrix.

We can apply this estimate to the matrix from the two-dimensional BVP, section 4.2.2.2.

Exercise 5.21. Show that in equation (4.14) the original matrix has $O(N) = O(n^2)$ nonzero elements, $O(N^2) = O(n^4)$ elements in total, and the factorization has $O(nN) = O(n^3) = O(N^{3/2})$ nonzeros.

These estimates show that the storage required for an LU factorization can be more than what is required for A, and the difference is not a constant factor, but related to the matrix size. Without proof we state that the inverses of the kind of sparse matrices you have seen so far are fully dense, so storing them takes even more. This is an important reason that solving linear systems $Ax = y$ is not done in practice by computing A^{-1} and multiplying $x = A^{-1}y$. (Numerical stability is another reason that this is not done.) The fact that even a factorization can take a lot of space is one reason for considering iterative methods, as we will do in section 5.5.

Above, you saw that the factorization of a dense matrix of size $n \times n$ takes $O(n^3)$ operations. How is this for a sparse matrix? Let us consider the case of a matrix with halfbandwidth p, and assume that the original matrix is dense in that band. The pivot element a_{11} is used to zero p elements in the first column, and for each the first row is added to that row, involving p multiplications and additions. In sum, we find that the number of operations is

roughly

$$\sum_{i=1}^{n} p^2 = p^2 \cdot n$$

plus or minus lower order terms.

Exercise 5.22. The assumption of a band that is initially dense is not true for the matrix of a two-dimensional BVP. Why does the above estimate still hold, up to some lower order terms?

5.4.3.2 Graph theory of sparse LU factorization

Graph theory is often useful when discussion the factorization of a sparse matrix. Let us investigate what eliminating the first unknown (or sweeping the first column) means in graph theoretic terms. We are assuming a structurally symmetric matrix.

We consider eliminating the first unknown as a process that takes a graph $G = \langle V, E\rangle$ and turns it into a graph $G' = \langle V', E'\rangle$. The relation between these graphs is first that the unknown has been removed from the vertices: $1 \notin V', V' \cup 1 = V$.

The relationship between E and E' is more complicated. In the Gaussian elimination algorithm the result of eliminating variable 1 is that the statement

$$a_{jk} \leftarrow a_{jk} - a_{a1}a_{11}^{-1}a_{1k}$$

is executed for all $j, k > 1$. If $a_{jk} \neq 0$ originally, then its value is merely altered. In case $a_{jk} = 0$ in the original matrix, there will be a nonzero fill-in element after the first unknown is eliminated. This means that in E there was no edge (j, k), and this edge *is* present in E'.

Summarizing, eliminating an unknown gives a graph that has one vertex less, and that has edges for all j, k such that there were edges between j or k and the eliminated variable.

Exercise 5.23. Prove the generalization of this. Let $I \subset V$ be any set of vertices, and let J be the vertices connected to I:

$$J \cap I = \emptyset, \quad \forall_{i\in I, j\in J} \colon (i, j) \in E.$$

Now show that eliminating the variables in I leads to a graph $\langle V', E'\rangle$ where all nodes in J are connected:

$$\forall_{j_1, j_2 \in J} \colon (j_1, j_2) \in E'.$$

5.4.3.3 Fill-in reduction

Graph properties of a matrix, such as degree and diameter, are invariant under renumbering the variables. Other properties, such as fill-in during a factorization, are affected by renumbering. In fact, it is worthwhile investigating whether it is possible to reduce the amount of fill-in by renumbering the nodes of the matrix graph, or equivalently, by applying a *permutation* to the linear system.

Exercise 5.24. Consider the 'arrow' matrix with nonzeroes only in the first row and column and on the diagonal:

$$\begin{pmatrix} * & * & \cdots & * \\ * & * & & \emptyset \\ \vdots & & \ddots & \\ * & \emptyset & & * \end{pmatrix}$$

What is the number of nonzeros in the matrix, and in the factorization, assuming that no addition ever results in zero? Can you find a symmetric permutation of the variables of the problem such that the new matrix has no fill-in?

The above estimates can sometimes be improved upon by clever permuting of the matrix, but in general the statement holds that an LU factorization of a sparse matrix takes considerably more space than the matrix itself. This is one of the motivating factors for the iterative methods in the next section.

5.5 Iterative methods

Gaussian elimination, the use of an LU factorization, is a simple way to find the solution of a linear system, but as we saw above, in the sort of problems that come from discretized PDEs, it can create a lot of fill-in. In this section we will look at a completely different approach, where the solution of the system is found by a sequence of approximations.

The computational scheme looks, very roughly, like:

$$\begin{cases} \text{Choose any starting vector } x_0 \text{ and repeat for } i \geq 0\text{:} \\ x_{i+1} = Bx_i + c \\ \text{until some stoppping test is satisfied.} \end{cases}$$

The important feature here is that no systems are solved with the original coefficient matrix; instead, every iteration involves a matrix-vector multiplication or a solution of a much simpler system. Thus we have replaced a complicated operation, constructing an LU factorization and solving a system with it, by a repeated simpler and cheaper application. This makes iterative methods easier to code, and potentially more efficient.

Let us consider a simple example to motivate the precise definition of the iterative methods. Suppose we want to solve the system

$$\begin{pmatrix} 10 & 0 & 1 \\ 1/2 & 7 & 1 \\ 1 & 0 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 21 \\ 9 \\ 8 \end{pmatrix}$$

which has the solution $(2,1,1)$. Suppose you know (for example, from physical considerations) that solution components are roughly the same size. Observe the dominant size of the diagonal, then, to decide that

$$\begin{pmatrix} 10 & & \\ & 7 & \\ & & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 21 \\ 9 \\ 8 \end{pmatrix}$$

might be a good approximation. This has the solution $(2.1, 9/7, 8/6)$. Clearly, solving a system that only involves the diagonal of the original system is both easy to do, and, at least in this case, fairly accurate.

Another approximation to the original system would be to use the lower triangle. The system

$$\begin{pmatrix} 10 & & \\ 1/2 & 7 & \\ 1 & 0 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 21 \\ 9 \\ 8 \end{pmatrix}$$

has the solution $(2.1, 7.95/7, 5.9/6)$. Solving triangular systems is a bit more work than diagonal systems, but still a lot easier than computing an LU factorization. Also, we have not generated any fill-in in the process of finding this approximate solution.

Thus we see that there are easy to compute ways of getting reasonably close to the solution. Can we somehow repeat this trick?

Formulated a bit more abstractly, what we did was instead of solving $Ax = b$ we solved $L\tilde{x} = b$. Now define Δx as the distance from the true solution: $\tilde{x} = x + \Delta x$. This gives $A\Delta x = A\tilde{x} - b \equiv r$, where r is the *residual*. Next we solve again $L\widetilde{\Delta x} = r$ and update $\tilde{\tilde{x}} = \tilde{x} - \widetilde{\Delta x}$.

iteration	1	2	3
x_1	2.1000	2.0017	2.000028
x_2	1.1357	1.0023	1.000038
x_3	0.9833	0.9997	0.999995

In this case we get two decimals per iteration, which is not typical.

It is now clear why iterative methods can be attractive. Solving a system by Gaussian elimination takes $O(n^3)$ operations, as shown above. A single iteration in a scheme as the above takes $O(n^2)$ operations if the matrix is dense, and possibly as low as $O(n)$ for a sparse matrix. If the number of iterations is low, this makes iterative methods competitive.

Exercise 5.25. When comparing iterative and direct methods, the flop count is not the only relevant measure. Outline some issues relating to the efficiency of the code in both cases.

5.5.1 Abstract presentation

It is time to do a formal presentation of the iterative scheme of the above example. Suppose we want to solve $Ax = b$, and a direct solution is too expensive, but multiplying by A is feasible. Suppose furthermore that we have a matrix $K \approx A$ such that solving $Kx = b$ can be done cheaply.

Instead of solving $Ax = b$ we solve $Kx = b$, and define x_0 as the solution: $Kx_0 = b$. This leaves us with an error $e_0 = x_0 - x$, for which we have the equation $A(x_0 - e_0) = b$ or $Ae_0 = Ax_0 - b$. We call $r_0 \equiv Ax_0 - b$ the *residual*; the error then satisfies $Ae_0 = r_0$.

If we could solve the error from the equation $Ae_0 = r_0$, we would be done: the true solution is then found as $x = x_0 - e_0$. However, since solving with A was too expensive the last time, we can not do so this time either, so we determine the error correction approximately. We solve $K\tilde{e}_0 = r_0$ and set $x_1 := x_0 - \tilde{e}_0$; the story can now continue with $e_1 = x_1 - x$, $r_1 = Ax_1 - b$, $K\tilde{e}_1 = r_1$, $x_2 = x_1 - \tilde{e}_1$, et cetera.

The iteration scheme is then:

Let x_0 be given
For $i \geq 0$:
    let $r_i = Ax_i - b$
    compute e_i from $Ke_i = r_i$
    update $x_{i+1} = x_i - e_i$

The scheme we have analyzed here is called *stationary iteration*, where every updated is performed the same way, without any dependence on the iteration number. It has a simple analysis, but unfortunately limited applicability.

There are several questions we need to answer about iterative schemes:

- Does this scheme alway take us to the solution?
- If the scheme converges, how quickly?
- When do we stop iterating?
- How do we choose K?

We will now devote some attention to these matters, though a full discussion is beyond the scope of this book.

5.5.2 Convergence and error analysis

We start with the question of whether the iterative scheme converges, and how quickly. Consider one iteration step:

$$\begin{aligned} r_1 &= Ax_1 - b = A(x_0 - \tilde{e}_0) - b \\ &= r_0 - AK^{-1}r_0 \\ &= (I - AK^{-1})r_0 \end{aligned} \tag{5.9}$$

Inductively we find $r_n = (I - AK^{-1})^n r_0$, so $r_n \downarrow 0$ if $|\lambda(I - AK^{-1})| < 1$.

This last statement gives us both a condition for convergence, by relating K to A, and a geometric convergence rate, if K is close enough.

Exercise 5.26. Derive a similar inductive relation for e_n.

It is hard to determine if the condition $|\lambda(I - AK^{-1})| < 1$ is satisfied by computing the actual eigenvalues. However, sometimes the Gershgorin theorem (appendix A.1.3) gives us enough information.

Exercise 5.27. Consider the matrix A of equation (4.14) that we obtained from discretization of a two-dimensional BVP. Let K be matrix containing the diagonal of A, that is $k_{ii} = a_{ii}$ and $k_{ij} = 0$ for $i \neq j$. Use the Gershgorin theorem to show that $|\lambda(I - AK^{-1})| < 1$.

The argument in this exercise is hard to generalize for more complicated choices of K, such as you will see below. Here, we only remark that for certain matrices A, these choices of K will always lead to convergence, with a speed that decreases as the matrix size increases. We will not go into the details, beyond stating that for M-matrices (see section 4.2.2.1) these iterative methods converge. For more details on the convergence theory of stationary iterative methods, see [82]

5.5.3 Computational form

Above, in section 5.5.1, we derived stationary iteration as a process that involves multiplying by A and solving with K. However, in some cases a simpler implementation is possible. Consider the case where $A = K - N$, and we know both K and N. Then we write

$$Ax = b \Rightarrow Kx = Nx + b$$

and we observe that x is a fixed point of the iteration

$$Kx_{i+1} = Nx_i + b.$$

It is easy to see that this is a stationary iteration:

$$\begin{aligned} Kx_{i+1} &= Nx_i + b \\ &= Kx_i - Ax_i + b \\ &= Kx_i - r_i \\ \Rightarrow x_{i+1} &= x_i - K^{-1}r_i. \end{aligned}$$

The congence criterium $|\lambda(I - AK^{-1})| < 1$ (see section 5.5.2) now simplifies to $|\lambda(NK^{-1})| < 1$.

Let us consider some cases. First of all, let $K = D_A$, that is, the matrix containing the diagonal part of A: $k_{ii} = a_{ii}$ and $k_{ij} = 0$ for all $i \neq j$. This is known as the *Jacobi method* method. The iteration scheme becomes

for $k = 1, \ldots$ until convergence, do:
  for $i = 1 \ldots n$:
    $x_i^{(k+1)} = a_{ii}^{-1}(\sum_{j \neq i} a_{ij} x_j^{(k)} + b_i)$

This requires us to have one vector x for the current iterate $x^{(k)}$, and one temporary t for the next vector $x^{(k+1)}$. The easiest way to write this is probably:

for $k = 1, \ldots$ until convergence, do:
  for $i = 1 \ldots n$:
    $t_i = a_{ii}^{-1}(-\sum_{j \neq i} a_{ij} x_j + b_i)$
  copy $x \leftarrow t$

But, you might think, in the sum $\sum_{j\neq i} a_{ij}x_j$ why not use the $x^{(k+1)}$ values for as far as already computed? In terms of the vectors $x^{(k)}$ this means

for $k = 1, \ldots$ until convergence, do:
  for $i = 1 \ldots n$:
    $x_i^{(k+1)} = a_{ii}^{-1}(-\sum_{j<i} a_{ij}x_j^{(k+1)}) - \sum_{j>i} a_{ij}x_j^{(k)} + b_i)$

Surprisingly, the implementation is simpler than of the Jacobi method:

for $k = 1, \ldots$ until convergence, do:
  for $i = 1 \ldots n$:
    $x_i = a_{ii}^{-1}(-\sum_{j\neq i} a_{ij}x_j + b_i)$

If you write this out as a matrix equation, you see that the newly computed elements elements $x_i^{(k+1)}$ are multiplied with elements of $D_A + L_A$, and the old elements $x_j^{(k)}$ by U_A, giving

$$(D_A + L_A)x^{(k+1)} = -U_A x^{(k)} + b$$

which is called the *Gauss-Seidel* method.

Finally, we can insert a damping parameter into the Gauss-Seidel scheme, giving the Successive Over-Relaxation (SOR) method:

for $k = 1, \ldots$ until convergence, do:
  for $i = 1 \ldots n$:
    $x_i^{(k+1)} = \omega a_{ii}^{-1}(-\sum_{j<i} a_{ij}x_j^{(k+1)}) - \sum_{j>i} a_{ij}x_j^{(k)} + b_i) + (1-\omega)x^{(k)}$

Surprisingly for something that looks like an interpolation, the method actually works with value for ω in the range $\omega \in (0, 2)$, the optimal value being larger than 1. Computing the optimal ω is not simple.

5.5.4 Convergence of the method

We are interested in two questions: firstly whether the iterative method converges at all, and if so, with what speed. The theory behind these questions goes far beyond this book. Above we remarked that convergence can often be guaranteed for M-matrices; with regard the convergence speed a full analysis is usually only possible in model cases. For the matrices from BVPs, as described in section 4.2.2.2, we state without proof that the smallest eigenvalue of the coefficient matrix is $O(h^2)$. The geometric convergence ratio $|\lambda(I - AK^{-1})|$ derived above can then be shown to be as follows:

- For the Jacobi method, the ratio is $1 - O(h^2)$;
- For the Gauss-Seidel iteration it also is $1 - O(h^2)$, but the method converges twice as fast;
- For the SOR method, the optimal omega can improve the convergence factor to $1 - O(h)$.

5.5.5 Choice of K

The convergence and error analysis above showed that the closer K is to A, the faster the convergence will be. In the initial examples we already saw the diagonal and lower triangular choice for K. We can describe these formally by letting $A = D_A + L_A + U_A$ be a splitting into diagonal, lower triangular, upper triangular part of A. Here are some methods with their traditional names:

- Richardson iteration: $K = \alpha I$.
- Jacobi method: $K = D_A$ (diagonal part),
- Gauss-Seidel method: $K = D_A + L_A$ (lower triangle, including diagonal)
- The SOR method: $K = \omega^{-1} D_A + L_A$
- Symmetric SOR (SSOR) method: $K = (D_A + L_A) D_A^{-1} (D_A + U_A)$.
- Iterative refinement: $K = LU$ where LU is a true factorization of A. In exact arithmetic, solving a system $LUx = y$ gives you the exact solution, so using $K = LU$ in an iterative method would give convergence after one step. In practice, roundoff error will make the solution be inexact, so people will sometimes iterate a few steps to get higher accuracy.

Exercise 5.28. What is the extra cost of a few steps of iterative refinement over a single system solution, assuming a dense system?

Exercise 5.29. The Jacobi iteration for the linear system $Ax = b$ is defined as

$$x_{i+1} = x_i - K^{-1}(Ax_i - b)$$

where K is the diagonal of A. Show that you can transform the linear system (that is, find a different coefficient matrix and right hand side vector) so that you can compute the same x_i vectors but with $K = I$, the identity matrix.
What are the implications of this strategy, in terms of storage and operation counts? Are there special implications if A is a sparse matrix?
Suppose A is symmetric. Give a simple example to show that $K^{-1}A$ does not have to be symmetric. Can you come up with a different transformation of the system so that symmetry is preserved and that has the same advantages as the transformation above? You can assume that the matrix has positive diagonal elements.

Exercise 5.30. Show that the transformation of the previous exercise can also be done for the Gauss-Seidel method. Give several reasons why this is not a good idea.

There are many different ways of choosing the preconditioner matrix K. Some of them are defined algebraically, such as the incomplete factorization discussed below. Other choices are inspired by the differential equation. For instance, if the operator is

$$\frac{\delta}{\delta x}(a(x,y)\frac{\delta}{\delta x}u(x,y)) + \frac{\delta}{\delta y}(b(x,y)\frac{\delta}{\delta y}u(x,y)) = f(x,y)$$

then the matrix K could be derived from the operator

$$\frac{\delta}{\delta x}(\tilde{a}(x)\frac{\delta}{\delta x}u(x,y)) + \frac{\delta}{\delta y}(\tilde{b}(y)\frac{\delta}{\delta y}u(x,y)) = f(x,y)$$

for some choices of $\tilde{a}, \tilde{b}$. The second set of equations is called a *separable problem*, and there are fast solvers; see [84].

5.5.5.1 Constructing K as an incomplete LU factorization

We briefly mention one other choice of K, which is inspired by Gaussian elimination. As in Gaussian elimination, we let $K = LU$, but now we use an *incomplete factorization*. Remember that a regular LU factorization is expensive because of the fill-in phenomenon. In an incomplete factorization, we limit the fill-in artificially.

If we write Gauss elimination as

```
for k,i,j:
   a[i,j] = a[i,j] - a[i,k] * a[k,j] / a[k,k]
```

we define an incomplete variant by

```
for k,i,j:
  if a[i,j] not zero:
    a[i,j] = a[i,j] - a[i,k] * a[k,j] / a[k,k]
```

- The resulting factorization is no longer exact: $LU \approx A$, so it is called an Incomplete LU (ILU) factorization.
- An ILU factorization takes much less space than a full factorization: the sparsity of $L + U$ the same as of A.

The algorithm above is called 'ILU(0)', where the zero refers to the fact that absolutely no fill-in is allowed during the incomplete factorization. Other schemes that allow a limited amount of fill-in exist. Much more can be said about this method; we will only remark that for M-matrices this scheme typically gives a converging method.

Exercise 5.31. How do operation counts of the matrix-vector product and solving a system with an ILU factorization compare?

You have now see that a full factorization of sparse matrix can need a higher order storage ($N^{3/2}$ for the factorization versus N for the matrix), but that an incomplete factorization takes $O(N)$, just like the matrix. It may therefore come as a surprise that the error matrix $R = A - LU$ is not dense, but itself sparse.

Exercise 5.32. Let A be the matrix of the Poisson equation, LU an incomplete factorization, and $R = A - LU$. Show that R is a bi-diagonal matrix:

- Consider that R consists of those elements that are discarded during the factorization. Where are they located in the matrix?
- Alternatively, write out the sparsity pattern of the product LU and compare that to the sparsity pattern of A.

5.5.5.2 Cost of constructing a preconditioner

In the example of the heat equation (section 4.3) you saw that each time step involves solving a linear system. As an important practical consequence, any setup cost for solving the linear system, such as constructing the preconditioner, will be amortized over the sequence of systems that is to be solved. A similar argument holds in the context of

nonlinear equations, a topic that we will not discuss as such. Nonlinear equations are solved by an iterative process such as the *Newton method*, which in its multidimensional form leads to a sequence of linear systems. Although these have different coefficient matrices, it is again possible to amortize setup costs.

5.5.6 Stopping tests

The next question we need to tackle is when to stop iterating. Above we saw that the error decreases geometrically, so clearly we will never reach the solution exactly, even if that were possible in computer arithmetic. Since we only have this relative convergence behaviour, how do we know when we are close enough?

We would like the error $e_i = x - x_i$ to be small, but measuring this is impossible. Above we observed that $Ae_i = r_i$, so

$$\|e_i\| \leq \|A^{-1}\|\|r_i\| \leq \lambda_{\max}(A^{-1})\|r_i\|$$

If we know anything about the eigenvalues of A, this gives us a bound on the error. (The norm of A is only the largest eigenvalue for symmetric A. In general, we need singular values here.)

Another possibility is to monitor changes in the computed solution. If these are small:

$$\|x_{n+1} - x_n\|/\|x_n\| < \epsilon$$

we can also conclude that we are close to the solution.

Exercise 5.33. Prove an analytic relationship between the distance between iterates and the distance to the true solution. If your equation contains constants, can they be determined theoretically or in practice?

Exercise 5.34. Write a simple program to experiment with linear system solving. Take the matrix from the 1D BVP (use an efficient storage scheme) and program an iterative method using the choice $K = D_A$. Experiment with stopping tests on the residual and the distance between iterates. How does the number of iterations depend on the size of the matrix?
Change the matrix construction so that a certain quantity is added the diagonal, that is, add αI to the original matrix. What happens when $\alpha > 0$? What happens when $\alpha < 0$? Can you find the value where the behaviour changes? Does that value depend on the matrix size?

5.5.7 Theory of general iterative methods

Above, you saw iterative methods of the form $x_{i+1} = x_i - K^{-1}r_i$, and we will now see iterative methods of the more general form

$$x_{i+1} = x_i + \sum_{j \leq i} K^{-1} r_j \alpha_{ji}. \tag{5.10}$$

One might ask, 'why not introduce an extra parameter and write $x_{i+1} = \alpha_i x_i + \cdots$?' Here we give a short argument that the former scheme describes a large class of methods. Indeed, the current author is not aware of methods that fall outside this scheme.

We defined the residual, given an approximate solution $\tilde{x}$, as $\tilde{r} = A\tilde{x} - b$. For this general discussion we precondition the system as $K^{-1}Ax = K^{-1}b$. (See section 5.5.5 where we discussed transforming the linear system.) The corresponding residual for the initial guess $\tilde{x}$ is

$$\tilde{r} = K^{-1}A\tilde{x} - K^{-1}b.$$

We now find that

$$x = A^{-1}b = \tilde{x} - A^{-1}K\tilde{r} = \tilde{x} - (K^{-1}A)^{-1}\tilde{r}.$$

Now, the *Cayley-Hamilton theorem* states that for every A there exists a polynomial $\phi(x)$ such that

$$\phi(A) = 0.$$

We observe that we can write this polynomial ϕ as

$$\phi(x) = 1 + x\pi(x)$$

where π is another polynomial. Applying this to $K^{-1}A$, we have

$$0 = \phi(K^{-1}A) = I + K^{-1}A\pi(K^{-1}A) \Rightarrow (K^{-1}A)^{-1} = -\pi(K^{-1}A)$$

so that $x = \tilde{x} + \pi(K^{-1}A)\tilde{r}$. Now, if we let $x_0 = \tilde{x}$, then $\tilde{r} = K^{-1}r_0$, giving the equation

$$x = x_0 + \pi(K^{-1}A)K^{-1}r_0.$$

This equation suggests an iterative scheme: if we can find a series of polynomials $\pi^{(i)}$ of degree i to approximate π, it will give us a sequence of iterates

$$x_{i+1} = x_0 + \pi^{(i)}(K^{-1}A)K^{-1}r_0 = x_0 + K^{-1}\pi^{(i)}(AK^{-1})r_0 \tag{5.11}$$

that ultimately reaches the true solution. Multiplying this equation by A and subtracting b on both sides gives

$$r_{i+1} = r_0 + \tilde{\pi}^{(i)}(AK^{-1})r_0$$

where $\tilde{\pi}^{(i)}(x) = x\pi^{(i)}(x)$. This immediately gives us

$$r_i = \hat{\pi}^{(i)}(AK^{-1})r_0 \tag{5.12}$$

where $\hat{\pi}^{(i)}$ is a polynomial of degree i with $\hat{\pi}^{(i)}(0) = 1$. This statement can be used as the basis of a convergence theory of iterative methods. However, this goes beyond the scope of this book.

Let us look at a couple of instances of equation (5.12). For $i = 1$ we have

$$r_1 = (\alpha_1 AK^{-1} + \alpha_2 I)r_0 \Rightarrow AK^{-1}r_0 = \beta_1 r_1 + \beta_0 r_0$$

for some values α_i, β_i. For $i = 2$

$$r_2 = (\alpha_2(AK^{-1})^2 + \alpha_1 AK^{-1} + \alpha_0)r_0$$

for different values α_i. But we had already established that AK_0^{-1} is a combination of r_1, r_0, so now we have that

$$(AK^{-1})^2 r_0 \in [\![r_2, r_1, r_0]\!],$$

and it is clear how to show inductively that

$$(AK^{-1})^i r_0 \in [\![r_i, \ldots, r_0]\!]. \tag{5.13}$$

Substituting this in (5.11) we finally get

$$x_{i+1} = x_0 + \sum_{j \leq i} K^{-1} r_j \alpha_{ji}. \tag{5.14}$$

It is easy to see that the scheme (5.10) is of the form (5.14). With a little effort one can show that the reverse implication also holds.

Summarizing, the basis of iterative methods is a scheme where iterates get updated by all residuals computed so far:

$$x_{i+1} = x_i + \sum_{j \leq i} K^{-1} r_j \alpha_{ji}. \tag{5.15}$$

Compare that to the stationary iteration (section 5.5.1) where the iterates get updated from just the last residual, and with a coefficient that stays constant.

We can say more about the α_{ij} coefficients. If we multiply equation (5.15) by A, subtracting b from both sides, we find

$$r_{i+1} = r_i + \sum_{j \leq i} AK^{-1} r_j \alpha_{ji}. \tag{5.16}$$

Let us consider this equation for a moment. If we have a starting residual r_0, the next residual is computed as

$$r_1 = r_0 + AK^{-1} r_0 \alpha_{00}.$$

From this we get that $AK^{-1}r_0 = \alpha_{00}^{-1}(r_1 - r_0)$, so for the next residual,

$$\begin{aligned} r_2 &= r_1 + AK^{-1} r_1 \alpha_{11} + AK^{-1} r_0 \alpha_{01} \\ &= r_1 + AK^{-1} r_1 \alpha_{11} + \alpha_{00}^{-1}\alpha_{01}(r_1 - r_0) \\ \Rightarrow AK^{-1} r_1 &= \alpha_{11}^{-1}(r_2 - (1 + \alpha_{00}^{-1}\alpha_{01})r_1 + \alpha_{00}^{-1}\alpha_{01} r_0) \end{aligned}$$

We see that we can express $AK^{-1}r_1$ as a sum $r_2\beta_2 + r_1\beta_1 + r_0\beta_0$, and that $\sum_i \beta_i = 0$.

Generalizing this, we find (with different α_{ij} than above)

$$\begin{aligned}
r_{i+1} &= r_i + AK^{-1}r_i\delta_i + \textstyle\sum_{j\leq i+1} r_j\alpha_{ji} \\
r_{i+1}(1-\alpha_{i+1,i}) &= AK^{-1}r_i\delta_i + r_i(1+\alpha_{ii}) + \textstyle\sum_{j<i} r_j\alpha_{ji} \\
r_{i+1}\alpha_{i+1,i} &= AK^{-1}r_i\delta_i + \textstyle\sum_{j\leq i} r_j\alpha_{ji} && \text{substituting } \begin{array}{l}\alpha_{ii} := 1+\alpha_{ii} \\ \alpha_{i+1,i} := 1-\alpha_{i+1,i}\end{array} \\
& && \text{note that } \alpha_{i+1,i} = \textstyle\sum_{j\leq i}\alpha_{ji} \\
r_{i+1}\alpha_{i+1,i}\delta_i^{-1} &= AK^{-1}r_i + \textstyle\sum_{j\leq i} r_j\alpha_{ji}\delta_i^{-1} \\
r_{i+1}\alpha_{i+1,i}\delta_i^{-1} &= AK^{-1}r_i + \textstyle\sum_{j\leq i} r_j\alpha_{ji}\delta_i^{-1} \\
r_{i+1}\gamma_{i+1,i} & AK^{-1}r_i + \textstyle\sum_{j\leq i} r_j\gamma_{ji} && \text{substituting } \gamma_{ij} = \alpha_{ij}\delta_j^{-1}
\end{aligned}$$

and we have that $\gamma_{i+1,i} = \sum_{j\leq i}\gamma_{ji}$.

We can take this last equation and write is as $AK^{-1}R = RH$ where

$$H = \begin{pmatrix} -\gamma_{11} & -\gamma_{12} & \cdots & & \\ \gamma_{21} & -\gamma_{22} & -\gamma_{23} & \cdots & \\ 0 & \gamma_{32} & -\gamma_{33} & -\gamma_{34} & \\ \emptyset & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

In this, H is a so-called *Hessenberg matrix*: it is upper triangular with a single lower subdiagonal. Also we note that the elements of H in each column sum to zero.

Because of the identity $\gamma_{i+1,i} = \sum_{j\leq i}\gamma_{ji}$ we can subtract b from both sides of the equation for r_{i+1} and 'divide out A', giving

$$x_{i+1}\gamma_{i+1,i} = K^{-1}r_i + \sum_{j\leq i} x_j\gamma_{ji}.$$

This gives us the general form for iterative methods:

$$\begin{cases} r_i = Ax_i - b \\ x_{i+1}\gamma_{i+1,i} = K^{-1}r_i + \sum_{j\leq i} x_j\gamma_{ji} \\ r_{i+1}\gamma_{i+1,i} = AK^{-1}r_i + \sum_{j\leq i} r_j\gamma_{ji} \end{cases} \qquad \text{where } \gamma_{i+1,i} = \textstyle\sum_{j\leq i}\gamma_{ji}. \tag{5.17}$$

This form holds for many iterative methods, including the stationary iterative methods you have seen above. In the next sections you will see how the γ_{ij} coefficients follow from orthogonality conditions on the residuals.

5.5.8 Iterating by orthogonalization

The stationary methods described above (section 5.5.1) have been around in some form or another for a long time: Gauss described some variant in a letter to a student. They were perfected in the thesis of Young [88] in 1950; the

final reference is probably the book by Varga [82]. These methods are little used these days, except in the specialized context of *multigrid smoothers*, a topic not discussed in this course.

At almost the same time, the field of methods based on orthogonalization was kicked off by two papers [59, 50], though it took a few decades for them to find wide applicability. (For further history, see [42].)

The basic idea is as follows:

> If you can make all your residuals orthogonal to each other, and the matrix is of dimension n, then after n iterations you have to have converged: it is not possible to have an $n+1$-st residual that is orthogonal and nonzero. Since a zero residual means that the corresponding iterate is the solution, we conclude that after n iterations we have the true solution in hand.

With the size of matrices that contemporary codes generate this reasoning is no longer relevant: it is not computationally realistic to iterate for n iterations. Moreover, roundoff will probably destroy any accuracy of the solution. However, it was later realized [72] that such methods *are* a realistic option in the case of symmetric positive definite (SPD) matrices. The reasoning is then:

> The sequence of residuals spans a series of subspaces of increasing dimension, and by orthogonalizing, the new residuals are projected on these spaces. This means that they will have decreasing sizes.

In this section you will see the basic idea of iterating by orthogonalization. The method presented here is only of theoretical interest; next you will see the Conjugate Gradients (CG) and Generalized Minimum Residual (GMRES) methods that are the basis of many real-life applications.

Let us now take the basic scheme (5.17) and orthogonalize the residuals. Instead of the normal inner product we use the K^{-1}-inner product:

$$(x,y)_{K^{-1}} = x^t K^{-1} y$$

and we will force residuals to be K^{-1}-orthogonal:

$$\forall_{i\neq j}\colon r_i \perp_{K^{-1}} r_j \Leftrightarrow \forall_{i\neq j}\colon r_i K^{-1} r_j = 0$$

This is known as the Full Orthogonalization Method (FOM) scheme:

Let r_0 be given
For $i \geq 0$:
    let $s \leftarrow K^{-1} r_i$
    let $t \leftarrow AK^{-1} r_i$
    for $j \leq i$:
        let γ_j be the coefficient so that $t - \gamma_j r_j \perp r_j$
    for $j \leq i$:
        form $s \leftarrow s - \gamma_j x_j$
        and $t \leftarrow t - \gamma_j r_j$
    let $x_{i+1} = (\sum_j \gamma_j)^{-1} s$, $r_{i+1} = (\sum_j \gamma_j)^{-1} t$.

You may recognize the *Gram-Schmidt* orthogonalization in this (see appendix A.1.1 for an explanation): in each iteration r_{i+1} is initially set to Ar_i, and orthogonalized against r_j with $j \leq i$.

We can use *modified Gram-Schmidt* by rewriting the algorithm as:

Let r_0 be given
For $i \geq 0$:
    let $s \leftarrow K^{-1}r_i$
    let $t \leftarrow AK^{-1}r_i$
    for $j \leq i$:
        let γ_j be the coefficient so that $t - \gamma_j r_j \perp r_j$
        form $s \leftarrow s - \gamma_j x_j$
        and $t \leftarrow t - \gamma_j r_j$
    let $x_{i+1} = (\sum_j \gamma_j)^{-1} s$, $r_{i+1} = (\sum_j \gamma_j)^{-1} t$.

These two version of the FOM algorithm are equivalent in exact arithmetic, but differ in practical circumstances in two ways:

- The modified Gramm-Schmidt method is more numerically stable;
- The unmodified method allows you to compute all inner products simultaneously. We discuss this below in section 6.6.

Even though the FOM algorithm is not used in practice, these computational considerations carry over to the GMRES method below.

5.5.9 Coupled recurrence form of iterative methods

Above, you saw the general equation (5.17) for generating iterates and search directions. This equation is often split as

- An update of the x_i iterate from a single *search direction*:

  $$x_{i+1} = x_i - \delta_i p_i,$$

  and
- A construction of the search direction from the residuals known so far:

  $$p_i = K^{-1}r_i + \sum_{j<i} \beta_{ij} K^{-1} r_j.$$

  It is not hard to see inductively that we can also define

  $$p_i = K^{-1}r_i + \sum_{j<i} \gamma_{ji} p_j,$$

  and this last form is the one that is used in practice.

The iteration dependent coefficients are typically chosen to let the residuals satisfy various orthogonality conditions. For instance, one can choose to let the method be defined by letting the residuals be orthogonal ($r_i^t r_j = 0$ if $i \neq j$), or A-orthogonal ($r_i^t A r_j = 0$ if $i \neq j$). Many more schemes exist. Such methods can converge much faster than stationary iteration, or converge for a wider range of matrix types. Below we will see two such methods; their analysis, however, is beyond the scope of this course.

5.5.10 The method of Conjugate Gradients

In this section, we will derive the *CG* method, which is a specific implementation of the FOM algorithm. In particular, it has pleasant computational properties in the case of a symmetric coefficient matrix A.

The CG method takes as its basic form the coupled recurrences formulation described above (section 5.5.1), and the coefficients are defined by demanding that the sequence of residuals $r_0, r_1, r_2, \ldots$ satisfy

$$r_i^t K^{-1} r_j = 0 \quad \text{if } i \neq j.$$

Unlike most expositions in the literature, we start by deriving the CG method for nonsymmetric systems, and then show how it simplifies in the symmetric case.

In order to derive the CG method, we start with the basic equations

$$\begin{cases} x_{i+1} = x_i - \delta_i p_i \\ r_{i+1} = r_i - \delta_i A p_i \\ p_i = K^{-1} r_i + \sum_{j<i} \gamma_{ji} p_j, \end{cases} \tag{5.18}$$

where the first and third equation were introduced above, and the second can be found by multiplying the first by A (check this!).

We will now derive the coefficients in this method by induction. In essence, we assume that we have current residual r_{cur}, a residuals to be computed r_{new}, and a collection of known residuals R_{old}. Rather than using subscripts 'old, cur, new', we use the following convention:

- x_1, r_1, p_1 are the current iterate, residual, and search direction. Note that the subscript 1 does not denote the iteration number here.
- x_2, r_2, p_2 are the iterate, residual, and search direction that we are about to compute. Again, the subscript does not equal the iteration number.
- X_0, R_0, P_0 are all previous iterates, residuals, and search directions bundled together in a block of vectors.

In terms of these quantities, the update equations are then

$$\begin{cases} x_2 = x_1 - \delta_1 p_1 \\ r_2 = r_1 - \delta_i A p_1 \\ p_2 = K^{-1} r_2 + \upsilon_{12} p_1 + P_0 u_{02} \end{cases} \tag{5.19}$$

where δ_1, υ_{12} are scalars, and u_{02} is a vector with length the number of iterations before the current. We now derive $\delta_1, \upsilon_{12}, u_{02}$ from the orthogonality of the residuals. To be specific, the residuals have to be orthogonal under the K^{-1} inner product: we want to have

$$r_2^t K^{-1} r_1 = 0, \qquad r_2^t K^{-1} R_0 = 0.$$

Combining these relations gives us, for instance,

$$\left.\begin{array}{l} r_1^t K^{-1} r_2 = 0 \\ r_2 = r_1 - \delta_i A K^{-1} p_1 \end{array}\right\} \Rightarrow \delta_1 = \frac{r_1^t r_1}{r_1^t A K^{-1} p_1}.$$

Finding υ_{12}, u_{02} is a little harder. For this, we start by summarizing the relations for the residuals and search directions in equation (5.18) in block form as

$$(R_0, r_1, r_2) \left(\begin{array}{ccccc|c|c} 1 & & & & & & \\ -1 & 1 & & & & & \\ & \ddots & \ddots & & & & \\ \hline & & & -1 & 1 & & \\ \hline & & & & -1 & 1 & \end{array}\right) = A(P_0, p_1, p_2)\,\mathrm{diag}(D_0, d_1, d_2)$$

$$(P_0, p_1, p_2) \begin{pmatrix} I - U_{00} & -u_{01} & -u_{02} \\ & 1 & -\upsilon_{12} \\ & & 1 \end{pmatrix} = K^{-1}(R_0, r_1, r_2)$$

or abbreviated $RJ = APD$, $P(I - U) = R$ where J is the matrix with identity diagonal and minus identity subdiagonal. We then observe that

- $R^t K^{-1} R$ is diagonal, expressing the orthogonality of the residuals.
- Combining that $R^t K^{-1} R$ is diagonal and $P(I - U) = R$ gives that $R^t P = R^t K^{-1} R (I - U)^{-1}$. We now reason that $(I - U)^{-1}$ is upper diagonal, so $R^t P$ is upper triangular. This tells us quantities such as $r_2^t p_1$ are zero.
- Combining the relations for R and P, we get first that

  $$R^t K^{-t} A P = R^t K^{-t} R J D^{-1}$$

  which tells us that $R^t K^{-t} A P$ is lower bidiagonal. Expanding R in this equation gives

  $$P^t A P = (I - U)^{-t} R^t R J D^{-1}.$$

  Here D and $R^t K^{-1} R$ are diagonal, and $(I-U)^{-t}$ and J are lower triangular, so $P^t A P$ is lower triangular.
- This tells us that $P_0^t A p_2 = 0$ and $p_1^t A p_2 = 0$.
- Taking the product of $P_0^t A$, $p_1^t A$ with the definition of p_2 in equation (5.19) gives

  $$u_{02} = -(P_0^t A P_0)^{-1} P_0^t A K^{-1} r_2, \qquad \upsilon_{12} = -(p_1^t A p_1)^{-1} p_1^t A K^{-1} r_2.$$

Compute $r^{(0)} = b - Ax^{(0)}$ for some initial guess $x^{(0)}$
for $i = 1, 2, \ldots$
  solve $Mz^{(i-1)} = r^{(i-1)}$
  $\rho_{i-1} = r^{(i-1)^T} z^{(i-1)}$
  if $i = 1$
    $p^{(1)} = z^{(0)}$
  else
    $\beta_{i-1} = \rho_{i-1}/\rho_{i-2}$
    $p^{(i)} = z^{(i-1)} + \beta_{i-1} p^{(i-1)}$
  endif
  $q^{(i)} = Ap^{(i)}$
  $\alpha_i = \rho_{i-1}/p^{(i)^T} q^{(i)}$
  $x^{(i)} = x^{(i-1)} + \alpha_i p^{(i)}$
  $r^{(i)} = r^{(i-1)} - \alpha_i q^{(i)}$
  check convergence; continue if necessary
end

Figure 5.2: The Preconditioned Conjugate Gradient Method

- If A is symmetric, P^tAP is lower triangular (see above) and symmetric, so it is in fact diagonal. Also, $R^tK^{-t}AP$ is lower bidiagonal, so, using $A = A^t$, $P^tAK^{-1}R$ is upper bidiagonal. Since $P^tAK^{-1}R = P^tAP(I - U)$, we conclude that $I - U$ is upper bidiagonal, so, *only in the symmetric case*, $u_{02} = 0$.

Some observations about this derivation.

- Strictly speaking we are only proving necessary relations here. It can be shown that these are sufficient too.
- There are different formulas that wind up computing the same vectors, in exact arithmetic. For instance, it is easy to derive that $p_1^t r_1 = r_1^t r_1$, so this can be substituted in the formulas just derived. The implementation of the CG method as it is typically implemented, is given in figure 5.2.
- In the k-th iteration, computing $P_0^t Ar_2$ (which is needed for u_{02}) takes k inner products. First of all, inner products are disadvantageous in a parallel context. Secondly, this requires us to store all search directions indefinitely. This second point implies that both work and storage go up with the number of iterations. Contrast this with the stationary iteration scheme, where storage was limited to the matrix and a few vectors, and work in each iteration was the same.
- The objections just raised disappear in the symmetric case. Since u_{02} is zero, the dependence on P_0 disappears, and only the dependence on p_1 remains. Thus, storage is constant, and the amount of work per iteration is constant. The number of inner products per iteration can be shown to be just two.

Exercise 5.35. Do a flop count of the various operations in one iteration of the CG method. Assume that A is the matrix of a five-point stencil and that the preconditioner M is an incomplete factorization of A (section 5.5.5.1). Let N be the matrix size.

5.5.11 Derivation from minimization

The above derivation of the CG method is not often found in the literature. The typical derivation starts with a minimization problem with a symmetric positive definite (SPD) matrix A:

$$\text{For which vector } x \text{ with } \|x\| = 1 \text{ is } f(x) = 1/2\, x^t A x - b^t x \text{ minimal?} \tag{5.20}$$

If we accept the fact that the function f has a minimum, which follows from the positive definiteness, we find the minimum by computing the derivative

$$f'(x) = Ax - b.$$

and asking where $f'(x) = 0$. And, presto, there we have the original linear system.

Exercise 5.36. Derive the derivative formula above. (Hint: write out the definition of derivative as $\lim_{h \downarrow 0} \ldots$.) Note that this requires A to be symmetric.

For the derivation of the iterative method, the basic form

$$x_{i+1} = x_i + p_i \delta_i$$

is posited, the value of

$$\delta_i = \frac{r_i^t p_i}{p_1^t A p_i}$$

is then derived as the one that minimizes the function f along the line $x_i + \delta p_i$:

$$\delta_i = \operatorname*{argmin}_{\delta} \|f(x_i + p_i \delta)\|$$

The construction of the search direction from the residuals follows by induction proof from the requirement that the residuals be orthogonal. For a typical proof, see [12].

5.5.12 GMRES

In the discussion of the CG method above, it was pointed out that orthogonality of the residuals requires storage of all residuals, and k inner products in the k'th iteration. Unfortunately, it can be proved that the work savings of the CG method can, for all practical purposes, not be found outside of SPD matrices [30].

The *GMRES* method is a popular implementation of such full orthogonalization schemes. In order to keep the computational costs within bounds, it is usually implemented as a *restarted* method. That is, only a certain number (say $k = 5$ or 20) of residuals is retained, and every k iterations the method is restarted. The code can be found in figure 5.3.

Other methods exist that do not have the storage demands of GMRES, for instance QMR [35] and BiCGstab [81]. Even though by the remark above these can not orthogonalize the residuals, they are still attractive in practice.

$x^{(0)}$ is an initial guess
for $j = 1, 2, ...$
  Solve r from $Mr = b - Ax^{(0)}$
  $v^{(1)} = r/\|r\|_2$
  $s := \|r\|_2 e_1$
  for $i = 1, 2, ..., m$
    Solve w from $Mw = Av^{(i)}$
    for $k = 1, ..., i$
      $h_{k,i} = (w, v^{(k)})$
      $w = w - h_{k,i}v^{(k)}$
    end
    $h_{i+1,i} = \|w\|_2$
    $v^{(i+1)} = w/h_{i+1,i}$
    apply $J_1, ..., J_{i-1}$ on $(h_{1,i}, ..., h_{i+1,i})$
    construct J_i, acting on ith and $(i + 1)$st component
    of $h_{.,i}$, such that $(i + 1)$st component of $J_i h_{.,i}$ is 0
    $s := J_i s$
    if $s(i + 1)$ is small enough then (UPDATE($\tilde{x}, i$) and quit)
  end
  UPDATE($\tilde{x}, m$)
end

In this scheme UPDATE($\tilde{x}, i$)
replaces the following computations:

Compute y as the solution of $Hy = \tilde{s}$, in which
the upper $i \times i$ triangular part of H has $h_{i,j}$ as
its elements (in least squares sense if H is singular),
$\tilde{s}$ represents the first i components of s
$\tilde{x} = x^{(0)} + y_1 v^{(1)} + y_2 v^{(2)} + ... + y_i v^{(i)}$
$s^{(i+1)} = \|b - A\tilde{x}\|_2$
if $\tilde{x}$ is an accurate enough approximation then quit
else $x^{(0)} = \tilde{x}$

Figure 5.3: The Preconditioned GMRES(m) Method

5.5.13 Complexity

The efficiency of Gaussian elimination was fairly easy to assess: factoring and solving a system takes, deterministically, $\frac{1}{3}n^3$ operations. For an iterative method, the operation count is the product of the number of operations per iteration times the number of iterations. While each individual iteration is easy to analyze, there is no good theory to predict the number of iterations. (In fact, an iterative method may not even converge to begin with.) Added to this is the fact that Gaussian elimination can be coded in such a way that there is considerable cache reuse, making the algorithm run at a fair percentage of the computer's peak speed. Iterative methods, on the other hand, are much slower on a flops per second basis.

All these considerations make the application of iterative methods to linear system solving somewhere in between a craft and a black art. In practice, people do considerable experimentation to decide whether an iterative method will pay off, and if so, which method is preferable.

5.6 Further Reading

The standard work on numerical linear algebra is Golub and Van Loan's *Matrix Computations* [43]. It covers algorithms, error analysis, and computational details. Heath's *Scientific Computing* covers the most common types of computations that arise in scientific computing; this book has many excellent exercises and practical projects.

Error analysis of computations in computer arithmetic is the focus of Wilkinson's classic *Rounding errors in Algebraic Processes* [86] and Higham's more recent *Accuracy and Stability of Numerical Algorithms* [51].

Algorithms such as the LU factorization can be coded in several ways that are mathematically equivalent, but that have different computational behaviour. This issue, in the context of dense matrices, is the focus of van de Geijn and Quintana's *The Science of Programming Matrix Computations* [80].

Chapter 6

High performance linear algebra

In this section we will discuss a number of issues pertaining to linear algebra on parallel computers. We will take a realistic view of this topic, assuming that the number of processors is finite, and that the problem data is always large, relative to the number of processors. We will also pay attention to the physical aspects of the communication network between the processors.

We will start off with a short section on the asymptotic analysis of parallelism, after which we will analyze various linear algebra operations, including iterative methods, and their behaviour in the presence of a network with finite bandwidth and finite connectivity. This chapter will conclude with various short remarks regarding complications in algorithms that arise due to parallel execution.

6.1 Asymptotics

If we ignore limitations such as that the number of processors has to be finite, or the physicalities of the interconnect between them, we can derive theoretical results on the limits of parallel computing. This section will give a brief introduction to such results, and discuss their connection to real life high performance computing.

Consider for instance the matrix-matrix multiplication $C = AB$, which takes $2N^3$ operations where N is the matrix size. Since there are no dependencies between the operations for the elements of C, we can perform them all in parallel. If we had N^2 processors, we could assign each to an (i, j) coordinate in C, and have it compute c_{ij} in $2N$ time. Thus, this parallel operation has efficiency 1, which is optimal.

Exercise 6.1. Show that this algorithm ignores some serious issues about memory usage:

- If the matrix is kept in shared memory, how many simultaneous reads from each memory locations are performed?
- If the processors have local storage, how much duplication is there of the matrix elements?

Adding N numbers $\{x_i\}_{i=1...N}$ can be performed in $\log_2 N$ time with $N/2$ processors. As a simple example, consider the sum of n numbers: $s = \sum_{i=1}^{n} a_i$. If we have $n/2$ processors we could compute:

1. Define $s_i^{(0)} = a_i$.

2. Iterate with $j = 1, \ldots, \log_2 n$:
3. Compute $n/2^j$ partial sums $s_i^{(j)} = s_{2i}^{(j-1)} + s_{2i+1}^{(j-1)}$

We see that the $n/2$ processors perform a total of n operations (as they should) in $\log_2 n$ time. The efficiency of this parallel scheme is $O(1/\log_2 n)$, a slowly decreasing function of n.

Exercise 6.2. Show that, with the scheme for parallel addition just outlined, you can multiply two matrices in $\log_2 N$ time with $N^3/2$ processors. What is the resulting efficiency?

It is now a legitimate theoretical question to ask

- If we had infinitely many processors, what is the lowest possible time complexity for matrix-matrix multiplication, or
- Are there faster algorithms that still have $O(1)$ efficiency?

Such questions have been researched (see for instance [48]), but they have little bearing on high performance computing.

A first objection to these kinds of theoretical bounds is that they implicitly assume some form of shared memory. In fact, the formal model for these algorithms is called a *Programmable Random Access Machine (PRAM)*, where the assumption is that every memory location is accessible to any processor. Often an additional assumption is made that multiple access to the same location are in fact possible[1]. These assumptions are unrealistic in practice, especially in the context of scaling up the problem size and the number of processors.

But even if we take distributed memory into account, theoretical results can still be unrealistic. The above summation algorithm can indeed work unchanged in distributed memory, except that we have to worry about the distance between active processors increasing as we iterate further. If the processors are connected by a linear array, the number of 'hops' between active processors doubles, and with that, asymptotically, the computation time of the iteration. The total execution time then becomes $n/2$, a disappointing result given that we throw so many processors at the problem.

What if the processors are connected with a hypercube topology? It is not hard to see that the summation algorithm can then indeed work in $\log_2 n$ time. However, as $n \to \infty$, can we build a sequence of hypercubes of n nodes and keep the communication time between two connected constant? Since communication time depends on latency, which partly depends on the length of the wires, we have to worry about the physical distance between nearest neighbours.

The crucial question here is whether the hypercube (an n-dimensional object) can be embedded in 3-dimensional space, while keeping the distance (measured in meters) constant between connected neighbours. It is easy to see that a 3-dimensional grid can be scaled up arbitrarily while maintaining a unit wire length, but the question is not clear for a hypercube. There, the length of the wires may have to increase as n grows, which runs afoul of the finite speed of electrons.

We sketch a proof (see [32] for more details) that, in our three dimensional world and with a finite speed of light, speedup is limited to $\sqrt[4]{n}$ for a problem on n processors, no matter the interconnect. The argument goes as follows. Consider an operation that involves collecting a final result on one processor. Assume that each processor takes

1. This notion can be made precise; for instance, one talks of a CREW-PRAM, for Concurrent Read, Exclusive Write PRAM.

a unit volume of space, produces one result per unit time, and can send one data item per unit time. Then, in an amount of time t, at most the processors in a ball with radius t, that is, $O(t^3)$ processors can contribute to the final result; all others are too far away. In time T, then, the number of operations that can contribute to the final result is $\int_0^T t^3 dt = O(T^4)$. This means that the maximum achievable speedup is the fourth root of the sequential time.

Finally, the question 'what if we had infinitely many processors' is not realistic as such, but we will allow it in the sense that we will ask the *weak scaling* question (section 2.7.2) 'what if we let the problem size and the number of processors grow proportional to each other'. This question is legitimate, since it corresponds to the very practical deliberation whether buying more processors will allow one to run larger problems, and if so, with what 'bang for the buck'.

6.2 Parallel dense matrix-vector product

In designing a parallel version of an algorithm, one often proceeds by making a *data decomposition* of the objects involved. In the case of a matrix-vector operations such as the product $y = Ax$, we have the choice of starting with a vector decomposition, and exploring its ramifications on how the matrix can be decomposed, or rather to start with the matrix, and deriving the vector decomposition from it. In this case, it seems natural to start with decomposing the matrix rather than the vector, since it will be most likely of larger computational significance. We now have two choices:

1. We make a one-dimensional decomposition of the matrix, splitting it in block rows or block columns, and assigning each of these – or groups of them – to a processor.
2. Alternatively, we can make a two-dimensional decomposition, assigning to each processor one or more general submatrices.

We start by considering the decomposition in block rows. Consider a processor p and the set I_p of indices of rows that it owns[2], and let $i \in I_p$ be a row that is assigned to this processor. The elements in row i are used in the operation

$$y_i = \sum_j a_{ij} x_j$$

We now reason:

- If processor p has all x_j values, the matrix-vector product can trivially be executed, and upon completion, the processor has the correct values y_j for $j \in I_p$.
- This means that every processor needs to have a copy of x, which is wasteful. Also it raises the question of data integrity: you need to make sure that each processor has the correct value of x.
- In certain practical applications (for instance iterative methods, as you have seen before), the output of the matrix-vector product is, directly or indirectly, the input for a next matrix-vector operation. This is certainly the case for the power method which computes $x, Ax, A^2x, \ldots$. Since our operation started with each processor having the whole of x, but ended with it owning only the local part of Ax, we have a mismatch.

2. For ease of exposition we will let I_p be a contiguous range of indices, but any general subset is allowed.

Input: Processor number p; the elements x_i with $i \in I_p$; matrix elements A_{ij} with $i \in I_p$.

Output: The elements y_i with $i \in I_p$

for $i \in I_p$ **do**
  $s \leftarrow 0$
  for $j \in I_p$ **do**
    $s \leftarrow s + a_{ij}x_j$
  end
  for $j \notin I_p$ **do**
    send x_j from the processor that owns it to the current one, then
    $s \leftarrow s + a_{ij}x_j$
  end
end

Procedure Naive Parallel MVP($A, x_{local}, y_{local}, p$)

Figure 6.1: A naïvely coded parallel matrix-vector product

- Maybe it is better to assume that each processor, at the start of the operation, has only the local part of x, that is, those x_i where $i \in I_p$, so that the start state and end state of the algorithm are the same. This means we have to change the algorithm to include some communication that allows each processor to obtain those values x_i where $i \notin I_p$.

Exercise 6.3. Go through a similar reasoning for the case where the matrix is decomposed in block columns. Describe the parallel algorithm in detail, like above, without giving pseudo code.

Let us now look at the communication in detail: we will consider a fixed processor p and consider the operations it performs and the communication that necessitates.

The matrix-vector product evaluates, on processor p, for each $i \in I_p$, the summation

$$y_i = \sum_j a_{ij}x_j.$$

If $j \in I_p$, the instruction $y_i = y_i + a_{aij}x_j$ is trivial; let us therefore consider only the case $j \notin I_p$. It would be nice if we could just write the statement

```
y(i) = y(i) + a(i,j)*x(j)
```

and some lower layer would automatically transfer `x(j)`, from whatever processor it is stored on, to a local register. (The PGAS languages (section 2.5.3) aim to do this, but their efficiency is far from guaranteed.) An implementation, based on this optimistic view of parallelism, is given in figure 6.1.

The immediate problem with such a 'local' approach is that too much communication will take place.

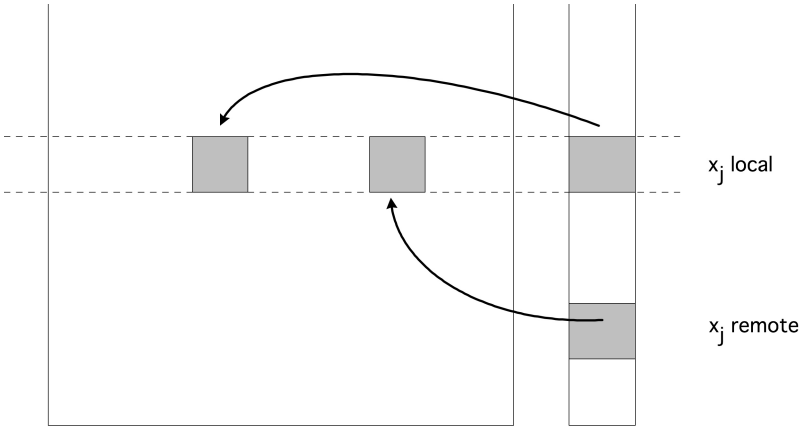


Figure 6.2: The parallel matrix-vector product with a blockrow distribution.

- If the matrix A is dense, the element x_j is necessary once for each row $i \in I_p$, and it will thus be fetched once for every row $i \in I_p$.
- For each processor $q \neq p$, there will be (large) number of elements x_j with $j \in I_q$ that need to be transferred from processor q to p. Doing this in separate messages, rather than one bulk transfer, is very wastefull.

With shared memory these issues are not much of a problem, but in the context of distributed memory it is better to take a *buffering* approach.

Instead of communicating individual elements of x, we use a local buffer B_{pq} for each processor $q \neq p$ where we collect the elements from q that are needed to perform the product on p. (See figure 6.2 for an illustration.) The parallel algorithm is given in figure 6.3.

In addition to preventing an element from being fetched more than once, this also combines many small messages into one large message, which is usually more efficient; recall our discussion of bandwidth and latency in section 2.6.6.

Exercise 6.4. Give pseudocode for the matrix-vector product using nonblocking operations (section 2.5.2.5)

Above we said that having a copy of the whole of x on each processor was wasteful in space. The implicit argument here is that, in general, we do not want local storage to be function of the number of processors: ideally it should be only a function of the local data. (This is related to weak scaling; section 2.7.2.)

Exercise 6.5. Make this precise. How many rows can we store locally, given a matrix size of N and local memory of M numbers? What is the exact buffer space required?

You see that, because of communication considerations, we have actually decided that it is unavoidable, or at least preferable, for each processor to store the whole input vector. Such trade-offs between space and time efficiency are fairly common in parallel programming. For the dense matrix-vector product we can actually defend this overhead, since the vector storage is of lower order than the matrix storage, so our over-allocation is small by ratio. Below, we will see that for the sparse matrix-vector product the overhead can be much less.

Input: Processor number p; the elements x_i with $i \in I_p$; matrix elements A_{ij} with $i \in I_p$.

Output: The elements y_i with $i \in I_p$

for $q \neq p$ **do**
Send elements of x from processor q to p, receive in buffer B_{pq}.

end

$y_{local} \leftarrow Ax_{local}$
for $q \neq p$ **do**
$y_{local} \leftarrow y_{local} + A_{pq}B_q$

end

Procedure Parallel MVP($A, x_{local}, y_{local}, p$)

Figure 6.3: A buffered implementation of the parallel matrix-vector product

It is easy to see that the parallel dense matrix-vector product, as described above, has perfect speedup *if we are allowed to ignore the time for communication*. In the next section you will see that the block row implementation above is not optimal if we take communication into account. For scalability (section 2.7.2) we need a two-dimensional decomposition.

6.3 Scalability of the dense matrix-vector product

In this section, we will give a full analysis of the parallel computation of $y \leftarrow Ax$, where $x, y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$. We will assume that p nodes will be used, but we make no assumptions on their connectivity. We will see that the way the matrix is distributed makes a big difference for the scaling of the algorithm; see section 2.7.2 for the defitions of the various forms of scaling.

An important part of this discussion is a thorough analysis of collective operations, so we start with this.

6.3.1 Collective operations

Collective operations play an important part in linear algebra operations. In fact, the scalability of the operations can depend on the cost of these collectives. (See [18] for details.)

In computing the cost of a collective operation, three architectural constants are enough to give lower bounds: α, the latency of sending a single message, β, the inverse of the bandwidth for sending data (see section 1.2.2), and γ, the time for performing an arithmetic operation. Sending n data items then takes time $\alpha + \beta n$. We further assume that a processor can only send one message at a time. We make no assumptions about the connectivity of the processors; thus, the lower bounds derived here will hold for a wide range of architectures.

The main implication of the architectural model above is that the number of active processors can only double in each step of an algorithm. For instance, to do a broadcast, first processor 0 sends to 1, then 0 and 1 can send to 2 and 3, then 0–3 send to 4–7, et cetera. This cascade of messages is called a *minimum spanning tree* of the processor network.

6.3.1.1 Broadcast

In a broadcast, a single processor sends its data to all others. By the above doubling argument, we conclude that a broadcast to p processors takes time at least $\lceil \log_2 p \rceil$ steps with a total latency of $\lceil \log_2 p \rceil \alpha$. Since n elements are sent, this adds $n\beta$, giving a total cost lower bound of

$$\lceil \log_2 p \rceil \alpha + n\beta.$$

We can illustrate the spanning tree method as follows:

	$t=0$	$t=1$	$t=2$	$t=3$
p_0	x_0, x_1, x_2, x_3	$x_0\downarrow, x_1\downarrow, x_2\downarrow, x_3\downarrow$	$x_0\downarrow, x_1\downarrow, x_2\downarrow, x_3\downarrow$	x_0, x_1, x_2, x_3
p_1			$x_0\downarrow, x_1\downarrow, x_2\downarrow, x_3\downarrow$	x_0, x_1, x_2, x_3
p_2				x_0, x_1, x_2, x_3
p_3				x_0, x_1, x_2, x_3

On $t=1$, p_0 sends to p_1; on $t=2$ p_0, p_1 send to p_2, p_3.

6.3.1.2 Reduction

By running the broadcast backwards in time, we see that a reduction operation has the same lower bound on the communication of $\lceil \log_2 p \rceil \alpha + n\beta$. A reduction operation also involves computation, with a total time of $(p-1)\gamma n$: each of n items gets reduced over p processors. Since these operations can potentially be parallelized, the lower bound on the computation is $\frac{p-1}{p}\gamma n$, giving a total of

$$\lceil \log_2 p \rceil \alpha + n\beta + \frac{p-1}{p}\gamma n.$$

We illustrate this again, using the notation $x_i^{(j)}$ for the data item i that was originally on processor j, and $x_i^{(j:k)}$ for the sum of the items i of processors $j \ldots k$.

	$t=0$	$t=1$	$t=2$	$t=3$
p_0	$x_0^{(0)}, x_1^{(0)}, x_2^{(0)}, x_3^{(0)}$	$x_0^{(0)}, x_1^{(0)}, x_2^{(0)}, x_3^{(0)}$	$x_0^{(0:1)}, x_1^{(0:1)}, x_2^{(0:1)}, x_3^{(0:1)}$	$x_0^{(0:3)}, x_1^{(0:3)}, x_2^{(0:3)}, x_3^{(0:3)}$
p_1	$x_0^{(1)}, x_1^{(1)}, x_2^{(1)}, x_3^{(1)}$	$x_0^{(1)}\uparrow, x_1^{(1)}\uparrow, x_2^{(1)}\uparrow, x_3^{(1)}\uparrow$		
p_2	$x_0^{(2)}, x_1^{(2)}, x_2^{(2)}, x_3^{(2)}$	$x_0^{(2)}, x_1^{(2)}, x_2^{(2)}, x_3^{(2)}$	$x_0^{(2:3)}\uparrow, x_1^{(2:3)}\uparrow, x_2^{(2:3)}\uparrow, x_3^{(2:3)}\uparrow$	
p_3	$x_0^{(3)}, x_1^{(3)}, x_2^{(3)}, x_3^{(3)}$	$x_0^{(3)}\uparrow, x_1^{(3)}\uparrow, x_2^{(3)}\uparrow, x_3^{(3)}\uparrow$		

On time $t=1$ processors p_0, p_2 receive from p_1, p_3, and on $t=2$ p_0 receives from p_2.

6.3.1.3 Allreduce

The cost of an allreduce is, somewhat remarkably, almost the same as of a simple reduction: since in a reduction not all processors are active at the same time, we assume that the extra work can be spread out perfectly. This means that the lower bound on the latency and computation stays the same. For the bandwidth we reason as follows: in order for the communication to be perfectly parallelized, $\frac{p-1}{p}n$ items have to arrive at, and leave each processor. Thus we have a total time of

$$\lceil \log_2 p \rceil \alpha + 2\frac{p-1}{p} n\beta + \frac{p-1}{p} n\gamma.$$

6.3.1.4 Allgather

Again we assume that gathers with multiple targets are active simultaneously. Since every processor originates a minimum spanning tree, we have $\log_2 p\alpha$ latency; since each processor receives n/p elements from $p-1$ processors, there is $\frac{p-1}{p}\beta$ latency. The total cost for constructing a length n vector is then

$$\lceil \log_2 p \rceil \alpha + \frac{p-1}{p} n\beta.$$

	$t=0$	$t=1$	$t=2$	$t=3$
p_0	x_0	$x_0\downarrow$	$x_0x_1\downarrow$	$x_0x_1x_2x_3$
p_1	x_1	$x_1\uparrow$	$x_0x_1\downarrow$	$x_0x_1x_2x_3$
p_2	x_2	$x_2\downarrow$	$x_2x_3\uparrow$	$x_0x_1x_2x_3$
p_3	x_3	$x_3\uparrow$	$x_2x_3\uparrow$	$x_0x_1x_2x_3$

At time $t=1$, there is an exchange between neighbours p_0, p_1 and likewise $p_2, p3$; at $t=2$ there is an exchange over distance two between p_0, p_2 and likewise p_1, p_3.

6.3.1.5 Reduce-scatter

In a *reduce-scatter* operations, processor i has an item $x_i^{(i)}$, and it needs $\sum_j x_i^{(j)}$. We could implement this by doing a size p reduction, collecting the vector $(\sum_i x_0^{(i)}, \sum_i x_1^{(i)}, \ldots)$ on one processor, and scattering the results. However it is possible to combine these operations:

	$t=0$	$t=1$	$t=2$	$t=3$
p_0	$x_0^{(0)}, x_1^{(0)}, x_2^{(0)}, x_3^{(0)}$	$x_0^{(0)}, x_1^{(0)}, x_2^{(0)}\downarrow, x_3^{(0)}\downarrow$	$x_0^{(0:2:2)}, x_1^{(0:2:2)}\downarrow$	$x_0^{(0:3)}$
p_1	$x_0^{(1)}, x_1^{(1)}, x_2^{(1)}, x_3^{(1)}$	$x_0^{(1)}, x_1^{(1)}, x_2^{(1)}\downarrow, x_3^{(1)}\downarrow$	$x_0^{(1:3:2)}\uparrow, x_1^{(1:3:2)}$	$x_1^{(0:3)}$
p_2	$x_0^{(2)}, x_1^{(2)}, x_2^{(2)}, x_3^{(2)}$	$x_0^{(2)}\uparrow, x_1^{(2)}\uparrow, x_2^{(2)}, x_3^{(2)}$	$x_2^{(0:2:2)}, x_3^{(0:2:2)}\downarrow$	$x_2^{(0:3)}$
p_3	$x_0^{(3)}, x_1^{(3)}, x_2^{(3)}, x_3^{(3)}$	$x_0^{(3)}\uparrow, x_1^{(3)}\uparrow, x_2^{(3)}, x_3^{(3)}$	$x_0^{(1:3:2)}\uparrow, x_1^{(1:3:2)}$	$x_3^{(0:3)}$

The reduce-scatter can be considered as a allgather run in reverse, with arithmetic added, so the cost is

$$\lceil \log_2 p \rceil \alpha + \frac{p-1}{p} n(\beta + \gamma).$$

6.3.2 Partitioning by rows

Partition

$$A \rightarrow \begin{pmatrix} A_0 \\ A_1 \\ \vdots \\ A_{p-1} \end{pmatrix} \quad x \rightarrow \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{p-1} \end{pmatrix}, \quad \text{and} \quad y \rightarrow \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{p-1} \end{pmatrix},$$

where $A_i \in \mathbb{R}^{m_i \times n}$ and $x_i, y_i \in \mathbb{R}^{m_i}$ with $\sum_{i=0}^{p-1} m_i = n$ and $m_i \approx n/p$. We will start by assuming that A_i, x_i, and y_i are originally assigned to $\mathcal{P}_i$.

The computation is characterized by the fact that each processor needs the whole vector x, but owns only an n/p fraction of it. Thus, we execute an *allgather* of x. After this, the processor can execute the local product $y_i \leftarrow A_i x$; no further communication is needed after that.

An algorithm with cost computation for $y = Ax$ in parallel is then given by

Step	Cost (lower bound)
Allgather x_i so that x is available on all nodes	$\lceil \log_2(p) \rceil \alpha + \frac{p-1}{p} n\beta \approx \log_2(p)\alpha + n\beta$
Locally compute $y_i = A_i x$	$\approx 2\frac{n^2}{p}\gamma$

Cost analysis The total cost of the algorithm is given by, approximately,

$$T_p(n) = T_p^{\text{1D-row}}(n) = 2\frac{n^2}{p}\gamma + \underbrace{\log_2(p)\alpha + n\beta}_{\text{Overhead}}.$$

Since the sequential cost is $T_1(n) = 2n^2\gamma$, the speedup is given by

$$S_p^{\text{1D-row}}(n) = \frac{T_1(n)}{T_p^{\text{1D-row}}(n)} = \frac{2n^2\gamma}{2\frac{n^2}{p}\gamma + \log_2(p)\alpha + n\beta} = \frac{p}{1 + \frac{p\log_2(p)}{2n^2}\frac{\alpha}{\gamma} + \frac{p}{2n}\frac{\beta}{\gamma}}$$

and the parallel efficiency by

$$E_p^{\text{1D-row}}(n) = \frac{S_p^{\text{1D-row}}(n)}{p} = \frac{1}{1 + \frac{p\log_2(p)}{2n^2}\frac{\alpha}{\gamma} + \frac{p}{2n}\frac{\beta}{\gamma}}.$$

An optimist's view Now, if one fixes p and lets n get large,

$$\lim_{n\rightarrow\infty} E_p(n) = \lim_{n\rightarrow\infty} \left[\frac{1}{1 + \frac{p\log_2(p)}{2n^2}\frac{\alpha}{\gamma} + \frac{p}{2n}\frac{\beta}{\gamma}}. \right] = 1.$$

Thus, if one can make the problem large enough, eventually the parallel efficiency is nearly perfect. However, this assumes unlimited memory, so this analysis is not practical.

A pessimist's view In a *strong scalability* analysis, one fixes n and lets p get large, to get

$$\lim_{p\rightarrow\infty} E_p(n) = \lim_{p\rightarrow\infty} \left[\frac{1}{1 + \frac{p\log_2(p)}{2n^2}\frac{\alpha}{\gamma} + \frac{p}{2n}\frac{\beta}{\gamma}}. \right] = 0.$$

Thus, eventually the parallel efficiency becomes nearly nonexistent.

A realist's view In a more realistic view we increase the number of processors with the amount of data. This is called *weak scalability*, and it makes the amount of memory that is available to store the problem scale linearly with p.

Let M equal the number of floating point numbers that can be stored in a single node's memory. Then the aggregate memory is given by Mp. Let $n_{\max}(p)$ equal the largest problem size that can be stored in the aggregate memory of p nodes. Then, if *all* memory can be used for the matrix,

$$(n_{\max}(p))^2 = Mp \quad \text{or} \quad n_{\max}(p) = \sqrt{Mp}.$$

The question now becomes what the parallel efficiency for the largest problem that can be stored on p nodes:

$$\begin{aligned} E_p^{\text{1D-row}}(n_{\max}(p)) &= \frac{1}{1 + \frac{p\log_2(p)}{2(n_{\max}(p))^2}\frac{\alpha}{\gamma} + \frac{p}{2n_{\max}(p)}\frac{\beta}{\gamma}} \\ &= \frac{1}{1 + \frac{\log_2(p)}{2M}\frac{\alpha}{\gamma} + \frac{\sqrt{p}}{2\sqrt{M}}\frac{\beta}{\gamma}}. \end{aligned}$$

Now, if one analyzes what happens when the number of nodes becomes large, one finds that

$$\lim_{p\rightarrow\infty} E_p(n_{\max}(p)) = \lim_{p\rightarrow\infty} \left[\frac{1}{1 + \frac{\log_2(p)}{2M}\frac{\alpha}{\gamma} + \frac{\sqrt{p}}{2\sqrt{M}}\frac{\beta}{\gamma}} \right] = 0.$$

Thus, this parallel algorithm for matrix-vector multiplication does not scale.

Alternatively, a realist realizes that he/she has a limited amount of time, $T_{\max}$ on his/her hands. Under the best of circumstances, that is, with zero communication overhead, the largest problem that we can solve in time $T_{\max}$ is given by

$$T_p(n_{\max}(p)) = 2\frac{(n_{\max}(p))^2}{p}\gamma = T_{\max}.$$

Thus

$$(n_{\max}(p))^2 = \frac{T_{\max}p}{2\gamma} \quad \text{or} \quad n_{\max}(p) = \frac{\sqrt{T_{\max}}\sqrt{p}}{\sqrt{2\gamma}}.$$

Then the parallel efficiency that is attained by the algorithm for the largest problem that can be solved in time $T_{\max}$ is given by

$$E_{p,n_{\max}} = \frac{1}{1 + \frac{\log_2 p}{T}\alpha + \sqrt{\frac{p}{T}\frac{\beta}{\gamma}}}$$

and the parallel efficiency as the number of nodes becomes large approaches

$$\lim_{p\to\infty} E_p = \sqrt{\frac{T\gamma}{p\beta}}.$$

Again, efficiency cannot be maintained as the number of processors increases and the execution time is capped.

6.3.3 Partitioning by columns

Partition

$$A \to (A_0, A_1, \ldots, A_{p-1}) \quad x \to \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{p-1} \end{pmatrix}, \quad \text{and} \quad y \to \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{p-1} \end{pmatrix},$$

where $A_j \in \mathbb{R}^{n\times n_j}$ and $x_j, y_j \in \mathbb{R}^{n_j}$ with $\sum_{j=0}^{p-1} n_j = n$ and $n_j \approx n/p$.

We will start by assuming that A_j, x_j, and y_j are originally assigned to $\mathcal{P}_j$ (but now A_i is a block of columns). In this algorithm by columns, processor i can compute the length n vector $A_i x_i$ without prior communication. These partial results then have to be added together

$$y \leftarrow \sum_i A_i x_i$$

in a *reduce-scatter* operation: each processor i scatters a part $(A_i x_i)_j$ of its result to processor j. The receiving processors then perform a reduction, adding all these fragments:

$$y_j = \sum_i (A_i x_i)_j.$$

The algorithm with costs is then given by:

Step	Cost (lower bound)
Locally compute $y^{(j)} = A_j x_j$	$\approx 2\frac{n^2}{p}\gamma$
Reduce-scatter the $y^{(j)}$s so that $y_i = \sum_{j=0}^{p-1} y_i^{(j)}$ is on $\mathcal{P}_i$	$\lceil \log_2(p) \rceil \alpha + \frac{p-1}{p} n\beta + \frac{p-1}{p} n\gamma \approx \log_2(p)\alpha + n(\beta+\gamma)$

Cost analysis The total cost of the algorithm is given by, approximately,

$$T_p^{\text{1D-col}}(n) = 2\frac{n^2}{p}\gamma + \underbrace{\log_2(p)\alpha + n(\beta+\gamma).}_{\text{Overhead}}$$

Notice that this is identical to the cost $T_p^{\text{1D-row}}(n)$, except with β replaced by $(\beta+\gamma)$. It is not hard to see that the conclusions about scalability are the same.

6.3.4 Two dimensional partitioning

Next, partition

$$A \rightarrow \begin{pmatrix} A_{00} & A_{01} & \cdots & A_{0,p-1} \\ A_{10} & A_{11} & \cdots & A_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ A_{p-1,0} & A_{p-1,0} & \cdots & A_{p-1,p-1} \end{pmatrix} \quad x \rightarrow \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{p-1} \end{pmatrix}, \quad \text{and} \quad y \rightarrow \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{p-1} \end{pmatrix},$$

where $A_{ij} \in \mathbb{R}^{n_i \times n_j}$ and $x_i, y_i \in \mathbb{R}^{n_i}$ with $\sum_{i=0}^{p-1} n_i = n$ and $n_i \approx n/p$.

We will view the nodes as an $r \times c$ mesh, with $p = rc$, and index them as p_{ij}, with $i = 0, \ldots, r-1$ and $j = 0, \ldots, c-1$. The following illustration for a 12×12 matrix on a 3×4 processor grid illustrates the assignment of data to nodes, where the i, j "cell" shows the matrix and vector elements owned by p_{ij}:

x_0				x_3				x_6				x_9			
a_{00}	a_{01}	a_{02}	y_0	a_{03}	a_{04}	a_{05}		a_{06}	a_{07}	a_{08}		a_{09}	$a_{0,10}$	$a_{0,11}$	
a_{10}	a_{11}	a_{12}		a_{13}	a_{14}	a_{15}	y_1	a_{16}	a_{17}	a_{18}		a_{19}	$a_{1,10}$	$a_{1,11}$	
a_{20}	a_{21}	a_{22}		a_{23}	a_{24}	a_{25}		a_{26}	a_{27}	a_{28}	y_2	a_{29}	$a_{2,10}$	$a_{2,11}$	
a_{30}	a_{31}	a_{32}		a_{33}	a_{34}	a_{35}		a_{37}	a_{37}	a_{38}		a_{39}	$a_{3,10}$	$a_{3,11}$	y_3
	x_1				x_4				x_7				x_{10}		
a_{40}	a_{41}	a_{42}	y_4	a_{43}	a_{44}	a_{45}		a_{46}	a_{47}	a_{48}		a_{49}	$a_{4,10}$	$a_{4,11}$	
a_{50}	a_{51}	a_{52}		a_{53}	a_{54}	a_{55}	y_5	a_{56}	a_{57}	a_{58}		a_{59}	$a_{5,10}$	$a_{5,11}$	
a_{60}	a_{61}	a_{62}		a_{63}	a_{64}	a_{65}		a_{66}	a_{67}	a_{68}	y_6	a_{69}	$a_{6,10}$	$a_{6,11}$	
a_{70}	a_{71}	a_{72}		a_{73}	a_{74}	a_{75}		a_{77}	a_{77}	a_{78}		a_{79}	$a_{7,10}$	$a_{7,11}$	y_7
		x_2				x_5				x_8				x_{11}	
a_{80}	a_{81}	a_{82}	y_8	a_{83}	a_{84}	a_{85}		a_{86}	a_{87}	a_{88}		a_{89}	$a_{8,10}$	$a_{8,11}$	
a_{90}	a_{91}	a_{92}		a_{93}	a_{94}	a_{95}	y_9	a_{96}	a_{97}	a_{98}		a_{99}	$a_{9,10}$	$a_{9,11}$	
$a_{10,0}$	$a_{10,1}$	$a_{10,2}$		$a_{10,3}$	$a_{10,4}$	$a_{10,5}$		$a_{10,6}$	$a_{10,7}$	$a_{10,8}$	y_{10}	$a_{10,9}$	$a_{10,10}$	$a_{10,11}$	
$a_{11,0}$	$a_{11,1}$	$a_{11,2}$		$a_{11,3}$	$a_{11,4}$	$a_{11,5}$		$a_{11,7}$	$a_{11,7}$	$a_{11,8}$		$a_{11,9}$	$a_{11,10}$	$a_{11,11}$	y_{11}

In other words, p_{ij} owns the matrix block A_{ij} and parts of x and y. This makes possible the following algorithm:

- Since x_j is distributed over the jth column, the algorithm starts by collecting x_j on each processor p_{ij} by an *allgather* inside the processor columns.
- Each processor p_{ij} then computes $y_{ij} = A_{ij}x_j$. This involves no further communication.
- The result y_i is then collected by gathering together the pieces y_{ij} in each processor row to form y_i, and this is then distributed over the processor row. These two operations are in fact combined to form a *reduce-scatter*.
- If $r = c$, we can transpose the y data over the processors, so that it can function as the input for a subsequent matrix-vector product. If, on the other hand, we are computing A^tAx, then y is now correctly distributed for the A^t product.

Algorithm The algorithm with cost analysis is

Step	Cost (lower bound)
Allgather x_i's within columns	$\lceil \log_2(r) \rceil \alpha + \frac{r-1}{p} n\beta \approx \log_2(r)\alpha + \frac{n}{c}\beta$
Perform local matrix-vector multiply	$\approx 2\frac{n^2}{p}\gamma$
Reduce-scatter y_i's within rows	$\lceil \log_2(c) \rceil \alpha + \frac{c-1}{p} n\beta + \frac{c-1}{p} n\gamma \approx \log_2(c)\alpha + \frac{n}{c}\beta + \frac{n}{c}\gamma$

Cost analysis The total cost of the algorithm is given by, approximately,

$$T_p^{r\times c}(n) = T_p^{r\times c}(n) = 2\frac{n^2}{p}\gamma + \underbrace{\log_2(p)\alpha + \left(\frac{n}{c} + \frac{n}{r}\right)\beta + \frac{n}{r}\gamma}_{\text{Overhead}}.$$

We will now make the simplification that $r = c = \sqrt{p}$ so that

$$T_p^{\sqrt{p}\times\sqrt{p}}(n) = T_p^{\sqrt{p}\times\sqrt{p}}(n) = 2\frac{n^2}{p}\gamma + \underbrace{\log_2(p)\alpha + \frac{n}{\sqrt{p}}(2\beta+\gamma)}_{\text{Overhead}}$$

Since the sequential cost is $T_1(n) = 2n^2\gamma$, the speedup is given by

$$S_p^{\sqrt{p}\times\sqrt{p}}(n) = \frac{T_1(n)}{T_p^{\sqrt{p}\times\sqrt{p}}(n)} = \frac{2n^2\gamma}{2\frac{n^2}{p}\gamma + \frac{n}{\sqrt{p}}(2\beta+\gamma)} = \frac{p}{1 + \frac{p\log_2(p)}{2n^2}\frac{\alpha}{\gamma} + \frac{\sqrt{p}}{2n}\frac{(2\beta+\gamma)}{\gamma}}$$

and the parallel efficiency by

$$E_p^{\sqrt{p}\times\sqrt{p}}(n) = \frac{1}{1 + \frac{p\log_2(p)}{2n^2}\frac{\alpha}{\gamma} + \frac{\sqrt{p}}{2n}\frac{(2\beta+\gamma)}{\gamma}}$$

We again ask the question what the parallel efficiency for the largest problem that can be stored on p nodes is.

$$\begin{aligned} E_p^{\sqrt{p}\times\sqrt{p}}(n_{\max}(p)) &= \frac{1}{1 + \frac{p\log_2(p)}{2n^2}\frac{\alpha}{\gamma} + \frac{\sqrt{p}}{2n}\frac{(2\beta+\gamma)}{\gamma}} \\ &= \frac{1}{1 + \frac{\log_2(p)}{2M}\frac{\alpha}{\gamma} + \frac{1}{2\sqrt{M}}\frac{(2\beta+\gamma)}{\gamma}} \end{aligned}$$

so that still

$$\lim_{p\rightarrow\infty} E_p^{\sqrt{p}\times\sqrt{p}}(n_{\max}(p)) = \lim_{p\rightarrow\infty} \frac{1}{1 + \frac{\log_2(p)}{2M}\frac{\alpha}{\gamma} + \frac{1}{2\sqrt{M}}\frac{(2\beta+\gamma)}{\gamma}} = 0.$$

However, $\log_2 p$ grows very slowly with p and is therefore considered to act much like a constant. In this case $E_p^{\sqrt{p}\times\sqrt{p}}(n_{\max}(p))$ decreases very slowly and the algorithm is considered to be scalable for practical purposes.

Note that when $r = p$ the 2D algorithm becomes the "partitioned by rows" algorithm and when $c = p$ it becomes the "partitioned by columns" algorithm. It is not hard to show that the 2D algorithm is scalable in the same sense as it is when $r = c$ as long as r/c is kept constant.

6.4 Scalability of LU factorization

A full analysis of the scalability of dense LU factorization is quite involved, so we will state without further proof that again a two-dimensional distribution is needed. However, we can identify a further complication. Since factorizations of any type[3] progress through a matrix, processors will be inactive for part of the time.

3. Gaussian elimination can be performed in right-looking, left-looking and something variants; see [80].

Exercise 6.6. Consider the regular right-looking Gaussian elimination

```
for k=1..n
  p = 1/a(k,k)
  for i=k+1,n
    for j=k+1,n
      a(i,j) = a(i,j)-a(i,k)*p*a(k,j)
```

Analyze the running time, speedup, and efficiency as a function of N, if we assume a one-dimensional distribution, and enough processors to store one column per processor. Show that speedup is limited.

Also perform this analysis for a two-dimensional decomposition where each processor stores one element.

For this reason, an *over-decomposition* is used, where the matrix is divided in more blocks than there are processors, and each processor stores several, non-contiguous, sub-matrices. Specifically, with $P < N$ processors, and assuming for simplicity $N = cP$, we let processor 0 store rows $0, c, 2c, 3c, \ldots$; processor 1 stores rows $1, c+1, 2c+1, \ldots$, et cetera. This scheme can be generalized two a two-dimensional distribution, if $N = c_1P_1 = c_2P_2$ and $P = P_1P_2$. This is called a 2D *cyclic distribution*. This scheme can be further extended by considering block rows and colums (with a small block size), and assigning to processor 0 the *block* rows $0, c, 2c, \ldots$.

Exercise 6.7. Consider a square $n \times n$ matrix, and a square $p \times p$ processor grid, where p divides n without remainder. Consider the over-decomposition outlined above, and make a sketch of matrix element assignment for the specific case $n = 6, p = 2$. That is, draw an $n \times n$ table where location (i, j) contains the processor number that stores the corresponding matrix element. Also make a table for each of the processors describing the local to global mapping, that is, giving the global (i, j) coordinates of the elements in the local matrix. (You will find this task facilitated by using zero-based numbering.)

Now write functions P, Q, I, J of i, j that describe the global to local mapping, that is, matrix element a_{ij} is stored in location $(I(i, j), J(i, j))$ on processor $(P(i, j), Q(i, j))$.

6.5 Parallel sparse matrix-vector product

The dense matrix-vector product, as you saw in the previous section, required each processor to communicate with every other, and to have a local buffer of essentially the size of the global vector. In the sparse case, considerably less buffer space is needed. Let us analyze this case.

The line $y_i = y_i + a_{ij}x_j$ now has to take into account that a_{ij} can be zero. In particular, we need to consider that, for some pairs $i \in I_p, j \notin I_p$ no communication will be needed. Declaring for each $i \in I_p$ a sparsity pattern set

$$S_{p;i} = \{j : j \notin I_p, a_{ij} \neq 0\}$$

our multiplication instruction becomes

$$y_i += a_{ij}x_j \qquad \text{if } j \in S_{p;i}.$$

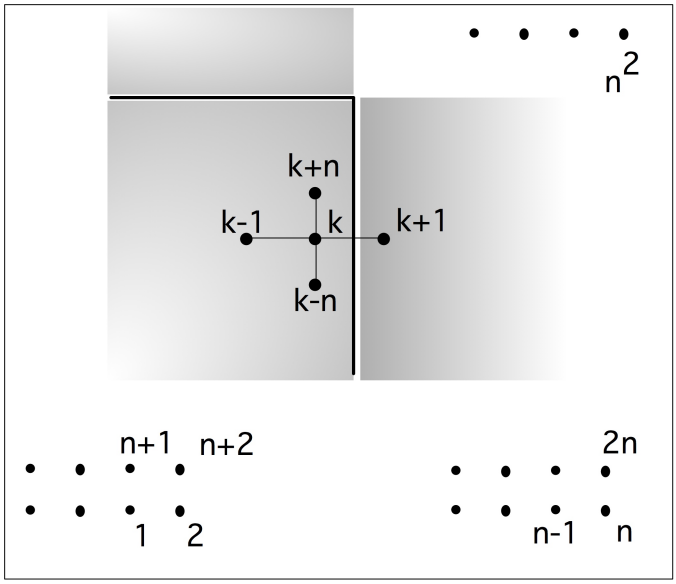


Figure 6.4: A difference stencil applied to a two-dimensional square domain, distributed over processors. A cross-processor connection is indicated.

If we want to avoid, as above, a flood of small messages, we combine all communication into a single message per processor. Defining

$$S_p = \cup_{i \in I_p} S_{p;i},$$

the algorithm now becomes:

- Collect all necessary off-processor elements x_j with $j \in S_p$ into one buffer;
- Perform the matrix-vector product, reading all elements of x from local storage.

This whole analysis of course also applies to dense matrices. This becomes different if we consider where sparse matrices come from. Let us start with a simple case.

Recall figure 4.1, which illustrated a discretized boundary value problem on the simplest domain, a square, and let us now parallelize it. We assume a natural ordering of the unknowns. A division of the matrix by blockrows corresponds to partitioning the domain over processors. Figure 6.4 shows how this gives rise to connections between processors: the elements a_{ij} with $i \in I_p, j \notin I_p$ are now the 'legs' of the stencil that reach beyond a processor boundary. The set of all such j, formally defined as

$$G = \{j \notin I_p : \exists_{i \in I_p} : a_{ij} \neq 0\}$$

is known as the *ghost region* of a processor; see figure 6.5.

Our crucial observation is now that, for each processor, the number of other processors it is involved with is strictly limited. This has implications for the efficiency of the operation.

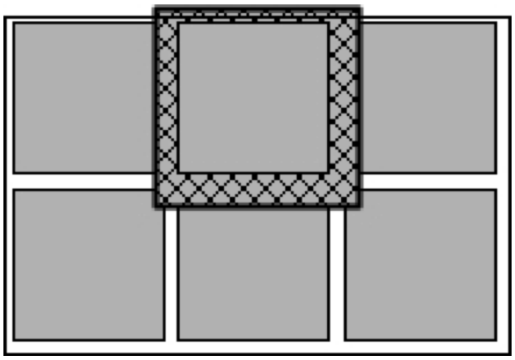


Figure 6.5: The ghost region of a processor, induced by a stencil

6.5.1 Parallel efficiency of the sparse matrix-vector product

In the case of the dense matrix-vector product (section 6.3), partitioning the matrix over the processors by (block) rows did not lead to a scalable algorithm. Part of the reason was the increase in the number of neighbours that each processors needs to communicate with. Figure 6.4 shows that, for the matrix of a 5-five point stencil, this number is limited to four.

Exercise 6.8. Take a square domain and a partitioning of the variables of the processors as in figure 6.4. What is the maximum number of neighbours a processor needs to communication with for the box stencil in figure 4.2? In three space dimensions, what is the number of neighbours if a 7-point central difference stencil is used? What can you say about the block structure of the matrix in each case?

The observation that each processor communicates with just a few neighbours stays intact if we go beyond square domains to more complicated physical objects. If a processor receives a more or less contiguous subdomain, the number of its neighbours will be limited. This implies that even in complicated problems each processor will only communicate with a small number of other processors. Compare this to the dense case where each processor had to receive data from *every* other processor. It is obvious that the sparse case is far more friendly to the interconnection network. (The fact that it also more common for large systems may influence the choice of network to install if you are about to buy a new parallel computer.)

For square domains, this argument can easily be made formal. Let the unit domain $[0,1]^2$ be partitioned over P processors in a $\sqrt{P} \times \sqrt{P}$ grid. Let the amount of work per processor be w and the communication time with each neighbour c. Then the time to perform the total work on a single processor is $T_1 = Pw$, and the parallel time is $T_P = w + 4c$, giving a speed up of

$$S_P = Pw/(w + 4c) = P/(1 + 4c/w) \approx P(1 - 4c/w).$$

Exercise 6.9. In this exercise you will analyze the parallel sparse matrix-vector product for a hypothetical, but realistic, parallel machine. Let the machine parameters be characterized by (see section 1.2.2):

- *Latency:* $\alpha = 1\mu s = 10^{-6}s$.
- *Bandwidth:* $1Gb/s$ corresponds to $\beta = 10^{-9}$.
- *Computation rate:* A per-core flops rate of $1G$flops means $\gamma = 10^9$. This number may seem low, but note that the matrix-vector product has less reuse than the matrix-matrix product, and that the sparse matrix-vector product is even more bandwidth-bound.

We assume 10^4 processors, and a five-point stencil matrix of size $N = 25 \cdot 10^{10}$. This means each processor stores $5 \cdot 8 \cdot N/p = 10^9$ bytes. If the matrix comes from a problem on a square domain, this means the domain was size $n \times n$ where $n = \sqrt{N} = 5 \cdot 10^5$.

Case 1. Rather than dividing the matrix, we divide the domain, and we do this first by horizontal slabs of size $n \times (n/p)$. Argue that the communication complexity is $2(\alpha + n\beta)$ and computation complexity is $10 \cdot n \cdot (n/p)$. Show that the resulting computation outweighs the communication by a factor 250.

Case 2. We divide the domain into patches of size $(n/\sqrt{p}) \times (n/\sqrt{p})$. The memory and computation time are the same as before. Derive the communication time and show that it is better by a factor of 50.

Argue that the first case does not weakly scale, that is, under the assumption that N/p is constant; the second case does scale.

The argument that a processor will only connect with a few neighbours is based on the nature of certain scientific computations. In the case of the *Boundary Element Method (BEM)*, any subdomain needs to communicate with everything in a radius r around it. As the number of processors goes up, the number of neighbours per processor will also go up.

Exercise 6.10. Give a formal analysis of the speedup and efficiency of the BEM algorithm. Assume again a unit amount of work w per processor and a time of communication c per neighbour. Since the notion of neighbour is now based on physical distance, not on graph properties, the number of neighbours will go up. Give T_1, T_p, S_p, E_p for this case.

There are also cases where a sparse matrix needs to be handled similarly to a dense matrix. For instance, *Google*'s *Pagerank* algorithm has at its heart the repeated operation $x \leftarrow Ax$ where A is a sparse matrix with $A_{ij} \neq 0$ if web page i links to j. This makes A a very sparse matrix, with no obvious structure, so every processor will most likely communicate with almost every other.

6.5.2 Setup of the sparse matrix-vector product

While in the square domain case it was easy for a processor to decide *who* its neighbours are, in the sparse case this is not so simple. The straightforward way to proceed is to have a preprocessing stage:

- Each processor makes an inventory of what non-local indices it needs; assuming that each processor knows what range of indices each other processor owns, it then decides which indices to get from what neighbours.
- Each processor sends a list of indices to each of its neighbours; this list will be empty for most of the neighbours, but we can not omit sending it.
- Each processor then receives these lists from all others, and draws up lists of which indices to send.

You will note that, even though the communication during the matrix-vector product involves only a few neighbours for each processor, giving a cost that is $O(1)$ in the number of processors, the setup involves all-to-all communications, which is $O(P)$ in the number of processors. The setup can be reduced to $O(\log P)$ with some trickery [31].

Exercise 6.11. The above algorithm for determining the communication part of the sparse matrix-vector product can be made far more efficient if we assume a matrix that is *structurally symmetric*: $a_{ij} \neq 0 \Leftrightarrow a_{ji} \neq 0$. Show that in this case no communication is needed to determine the communication pattern.

6.6 Computational aspects of iterative methods

All iterative methods feature the following operations:

- A matrix-vector product; this was discussed for the sequential case in section 5.4 and for the parallel case in section 6.5. In the parallel case, construction of FEM matrices has a complication that we will discuss in section 6.6.2.
- The construction of the preconditioner matrix $K \approx A$, and the solution of systems $Kx = y$. This was discussed in the sequential case in sections 5.5.5 and 5.5.5.1. Below we will go into parallelism aspects in section 6.7.2.
- Some vector operations (including inner products, in general). These will be discussed next.

6.6.1 Vector operations

Vector updates and inner products take only a small portion of the time of the algorithm.

Above we have already talked about the matrix-vector product.

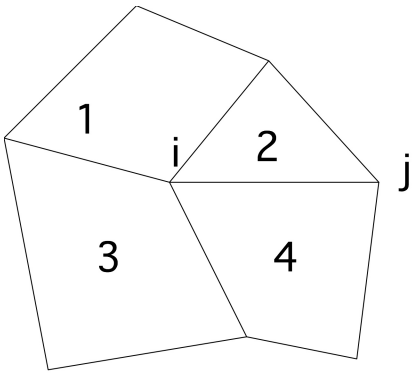
Vector updates are trivially parallel, so we do not worry about them. Inner products are more interesting. Since every processor is likely to need the value of the inner product, we use the following algorithm:

Algorithm: compute $a \leftarrow x^t y$ where x, y are distributed vectors.
For each processor p do:
compute $a_p \leftarrow x_p^t y_p$ where x_p, y_p are the part of x, y stored on processor p
do a global reduction to compute $a = \sum_p a_p$;
broadcast the result

The reduction and broadcast (which can be joined into an `Allgather`) combine data over all processors, so they have a communication time that increases with the number of processors. This makes the inner product potentially an expensive operation.

6.6.2 Finite element matrix construction

The *FEM* leads to an interesting issue in parallel computing. For this we need to sketch the basic outline of how this method works. The FEM derives its name from the fact that the physical objects modeled are divided into small two or three dimensional shapes such as triangles and squares in 2D, or pyramids and bricks in 3D. On each of these, the function we are modeling is then assumed to polynomial, often of a low degree, such as linear or bilinear.



The crucial fact is that a matrix element a_{ij} is then the sum of computations, specifically certain integrals, over all elements that variables i and j share:

$$a_{ij} = \sum_{e:\ i,j \in e} a_{ij}^{(e)}.$$

In the above figure, a_{ij} is the sum of computations over elements 2 and 4. Now, the computations in each element share many common parts, so it is natural to assign each element e uniquely to a processor P_e, which then computes all contributions $a_{ij}^{(e)}$.

In section 6.2 we described how each variable i was uniquely assigned to a processor P_i. Now we see that it is not possibly to make assignments P_e of elements and P_i of variables such that P_e computes in full the coefficients a_{ij} for all $i \in e$. In other words, if we compute the contributions locally, there needs to be some amount of communication to assemble certain matrix elements. For this reason, modern linear algebra libraries such as PETSc (see tutorial section B.8) allow any processor to set any matrix element.

6.7 Preconditioner construction, storage, and application

Above (sections 5.5.5 and 5.5.5.1) we saw a couple of different choices of K. Let us analyze their cost, in terms of storage and operations in the sequential case.

The Jacobi method used $K = D_A$, the diagonal of the matrix A. Thus it is not necessary to store K explicitly: we can reuse the elements of A. Solving the system $Kx = y$ for given y then takes n divisions. However, divisions can be quite expensive. It is a much better idea to store D_A^{-1} explicitly, and multiply by this matrix. Thus, we need as much storage as is needed for one vector, and the operation count is n multiplications per iteration.

The Gauss-Seidel method used $K = D_A + L_A$. A similar story holds as for Jacobi: theoretically we could get away with reusing the elements of A, but for efficiency we again store D_A^{-1} explicitly. Again, we need one vector's worth of storage; the operation count for solving $Kx = y$ is approximately half of a matrix vector multiplication.

Exercise 6.12. Instead of storing D_A^{-1}, we could also store $(D_A + L_A)^{-1}$. Give several reasons why this is a bad idea.

6.7.1 Construction of the preconditioner

This can be an expensive operations, taking the equivalent of several steps of the iterative method. If only one linear system is to be solved, this cost has to be weighed: a more expensive preconditioner leads to a faster converging method, but it is not *a priori* clear that the overall cost will be lower. However, often a number of systems need to be solved with the same coefficient matrix, for instance in an implicit scheme for an IBVP. In that case, the preconditioner construction cost is amortized over the linear systems.

6.7.2 Parallelism in the preconditioner

Above we saw that, in a flop counting sense, applying an ILU preconditioner (section 5.5.5.1) is about as expensive as doing a matrix-vector product. This is no longer true if we run our iterative methods on a parallel computer.

At first glance the operations are similar. A matrix-vector product $y = Ax$ looks like

```
for i=1..n
  y[i] = sum over j=1..n a[i,j]*x[j]
```

In parallel this would look like

```
for i=myfirstrow..mylastrow
  y[i] = sum over j=1..n a[i,j]*x[j]
```

Suppose that a processor has local copies of all the elements of A and x that it will need, then this operation is fully parallel: each processor can immediately start working, and if the work load is roughly equal, they will all finish at the same time. The total time for the matrix-vector product is then divided by the number of processors, making the speedup more or less perfect.

Consider now the forward solve $Lx = y$, for instance in the context of an ILU preconditioner:

```
for i=1..n
  x[i] = (y[i] - sum over j=1..i-1 ell[i,j]*x[j]) / a[i,i]
```

We can simply write the parallel code:

```
for i=myfirstrow..mylastrow
  x[i] = (y[i] - sum over j=1..i-1 ell[i,j]*x[j]) / a[i,i]
```

but now there is a problem. We can no longer say 'suppose a processor has local copies of everything in the right hand side', since the vector x appears both in the left and right hand side. While the matrix-vector product is in principle fully parallel over the matrix rows, this triangular solve code is recursive, hence sequential.

In a parallel computing context this means that, for the second processor to start, it needs to wait for certain components of x that the first processor computes. Apparently, the second processor can not start until the first one is finished, the third processor has to wait for the second, and so on. The disappointing conclusion is that in parallel only one processor will be active at any time, and the total time is the same as for the sequential algorithm. This is actually not a big problem in the dense matrix case, since parallelism can be found in the operations for handling a single row (see section 6.10), but in the sparse case it means we can not use incomplete factorizations without some redesign.

In the next few subsections we will see different strategies for finding preconditioners that perform efficiently in parallel.

6.7.2.1 Block Jacobi methods

Various approaches have been suggested to remedy this sequentiality the triangular solve. For instance, we could simply let the processors ignore the components of x that should come from other processors:

```
for i=myfirstrow..mylastrow
  x[i] = (y[i] - sum over j=myfirstrow..i-1 ell[i,j]*x[j]) / a[i,i]
```

This is not mathematically equivalent to the sequential algorithm (technically, it is called a *block Jacobi* method with ILU as the *local solve*), but since we're only looking for an approximationg $K \approx A$, this is simply a slightly cruder approximation.

Exercise 6.13. Take the Gauss-Seidel code you wrote above, and simulate a parallel run. What is the effect of increasing the (simulated) number of processors?

The idea behind block methods can easily be appreciated pictorially; see figure 6.6. In effect, we ignore all connec-

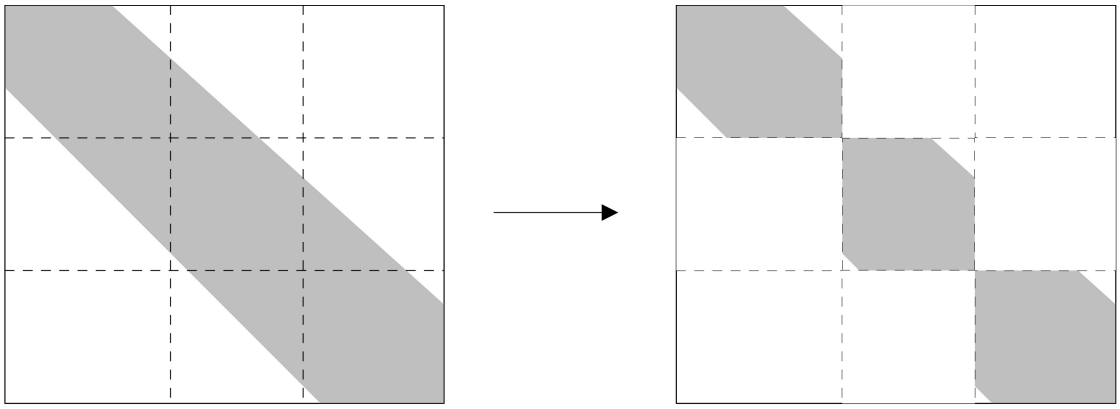


Figure 6.6: Sparsity pattern corresponding to a block Jacobi preconditioner

tions between processors. Since in a BVP all points influence each other (see section 4.2.1), using a less connected preconditioner will increase the number of iterations if executed on a sequential computer. However, block methods are parallel and, as we observed above, a sequential preconditioner is very inefficient in a parallel context.

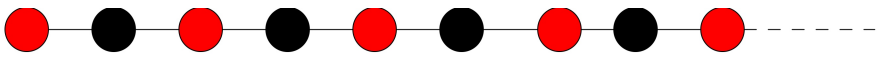


Figure 6.7: Red-black ordering of a the points on a line

6.7.2.2 *Variable reordering and colouring*

Another idea is to permute the matrix in a clever way. Let us take a simple example, where A is a tridiagonal matrix. The equation $Ax = b$ looks like

$$\begin{pmatrix} a_{11} & a_{12} & & & & \emptyset \\ a_{21} & a_{22} & a_{23} & & & \\ & a_{32} & a_{33} & a_{34} & & \\ & & \ddots & \ddots & \ddots & \\ \emptyset & & & & & \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{pmatrix}$$

We observe that x_i directly depends on x_{i-1} and x_{i+1}, but not x_{i-2} or x_{i+1}. Thus, let us see what happens if we permute the indices to group every other component together.

Pictorially, we take the points $1, \ldots, n$ and colour them red and black (figure 6.7), then we permute them to first take all red points, and subsequently all black ones. The correspondingly permuted matrix looks as follows:

$$\begin{pmatrix} a_{11} & & & a_{12} & & \\ & a_{33} & & a_{32} & a_{34} & \\ & & a_{55} & & \ddots & \ddots \\ & & & \ddots & & \\ a_{21} & a_{23} & & a_{22} & & \\ & a_{43} & a_{45} & & a_{44} & \\ & & \ddots & \ddots & & \ddots \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \\ x_5 \\ \vdots \\ x_2 \\ x_4 \\ \vdots \end{pmatrix} = \begin{pmatrix} y_1 \\ y_3 \\ y_5 \\ \vdots \\ y_2 \\ y_4 \\ \vdots \end{pmatrix}$$

With this permuted A, the Gauss-Seidel matrix $D_A + L_A$ looks like

$$\begin{pmatrix} a_{11} & & & \emptyset & & \\ & a_{33} & & & & \\ & & a_{55} & & & \\ & & & \ddots & & \\ a_{21} & a_{23} & & a_{22} & & \\ & a_{43} & a_{45} & & a_{44} & \\ & & \ddots & \ddots & & \ddots \end{pmatrix}$$

What does this buy us? Well, the odd numbered components $x_3, x_5, \ldots$ can now all be solved without any pre-requisites, so if we divide them over the processors, all processors will be active at the same time. After this, the even numbered components $x_2, x_4, \ldots$ can also be solved independently of each other, each needing the values of

the odd numbered solution components next to it. This is the only place where processors need to communicate: suppose a processor has $x_{100}, \ldots x_{149}$, then in the second solution stage the value of x_{99} needs to be sent from the previous processor.

Red-black ordering can be applied to two-dimensional problems too. Let us apply a red-black ordering to the points (i, j) where $1 \leq i, j \leq n$. Here we first apply a successive numbering to the odd points on the first line $(1, 1), (3, 1), (5, 1), \ldots$, then the even points of the second line $(2, 2), (4, 2), (6, 2), \ldots$, the odd points on the third line, et cetera. Having thus numbered half the points in the domain, we continue with the even points in the first line, the odd points in the second, et cetera. As you can see in figure 6.8, now the red points are only connected to

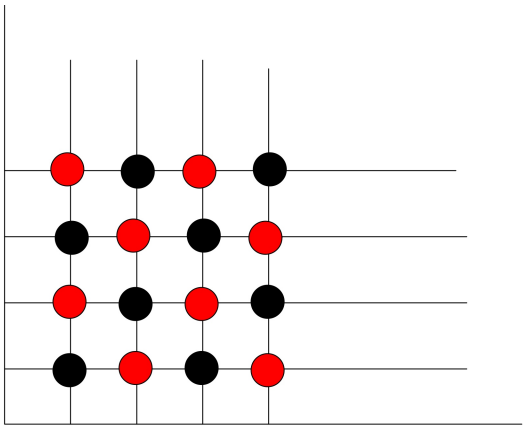


Figure 6.8: Red-black ordering of the variables of a two-dimensional domain

black points, and the other way around. In graph theoretical terms, you have found a *colouring* (see appendix A.6 for the definition of this concept) of the matrix graph with two colours.

Exercise 6.14. Apply the red-black ordering to the 2D BVP (4.10). Sketch the resulting matrix structure.

The red-black ordering is a simple example of *graph colouring* (sometimes called *multi-colouring*). In simple cases, such as the unit square domain we considered in section 4.2.2.2 or its extension to 3D, the *colour number* of the adjacency graph is easily determined.

Exercise 6.15. You saw that a red-black ordering of unknowns coupled with the regular five-point star stencil give two subsets of variables that are not connected among themselves, that is, they form a two-colouring of the matrix graph. Can you find a colouring if nodes are connected by the second stencil in figure 4.2?

There is a simple bound for the number of colours needed for the graph of a sparse matrix: the number of colours is at most $d + 1$ where d is the degree of the graph. To see that we can colour a graph with degree d using $d + 1$ colours, consider a node with degree d. No matter how its neighbours are coloured, there is always an unused colour among the $d + 1$ available ones.

Exercise 6.16. Consider a sparse matrix, where the graph can be coloured with d colours. Permute the matrix by first enumerating the unknowns of the first colour, then the second colour, et cetera. What can you say about the sparsity pattern of the resulting permuted matrix?

6.8 Trouble both ways

In some contexts, it is necessary to perform implicit calculations through all directions of a two or three-dimensional array. For example, in section 4.3 you saw how the implicit solution of the heat equation gave rise to repeated systems

$$(\alpha I + \frac{d^2}{dx^2} + \frac{d^2}{dy^2})u^{(t+1)} = u^{(t)} \tag{6.1}$$

Without proof, we state that the time-dependent problem can also be solved by

$$(\beta I + \frac{d^2}{dx^2})(\beta I + \frac{d^2}{dy^2})u^{(t+1)} = u^{(t)} \tag{6.2}$$

for suitable β. This scheme will not compute the same values on each individual time step, but it will converge to the same steady state.

This approach has considerable advantages, mostly in terms of operation counts: the original system has to be solved either making a factorization of the matrix, which incurs fill-in, or by solving it iteratively.

Exercise 6.17. Analyze the relative merits of these approaches, giving rough operation counts. Consider both the case where α has dependence on t and where it does not. Also discuss the expected speed of various operations.

A further advantage appears when we consider the parallel solution of (6.2). Note that we have a two-dimensional set of variables u_{ij}, but the operator $I+d^2u/dx^2$ only connects $u_{ij}, u_{ij-1}, u_{ij+1}$. That is, each line corresponding to an i value can be processed independently. Thus, both operators can be solved fully parallel using a one-dimensional partition on the domain. The solution of a the system in (6.1), on the other hand, has limited parallelism.

Unfortunately, there is a serious complication: the operator in x direction needs a partitioning of the domain in on direction, and the operator in y in the other. The solution usually taken is to transpose the u_{ij} value matrix in between the two solves, so that the same processor decomposition can handle both. This transposition can take a substantial amount of the processing time of each time step.

Exercise 6.18. Discuss the merits of and problems with a two-dimensional decomposition of the domain, using a grid of $P = p \times p$ processors. Can you suggest a way to ameliorate the problems?

One way to speed up these calculations, is to replace the implicit solve, by an explicit operation; see section 6.9.3.

6.9 Parallelism and implicit operations

In section 4.1.2.2 you saw that implicit operations can have great advantages from the point of numerical stability. However, you also saw that they make the difference between methods based on a simple operation such as the matrix-vector product, and ones based on the more complicated linear system solution. There are further problems with implicit methods when you start computing in parallel.

Exercise 6.19. Let A be the matrix

$$A = \begin{pmatrix} a_{11} & & \emptyset & \\ a_{21} & a_{22} & & \\ & \ddots & \ddots & \\ \emptyset & & a_{n,n-1} & a_{nn} \end{pmatrix}. \tag{6.3}$$

Show that the matrix vector product $y \leftarrow Ax$ and the system solution $x \leftarrow A^{-1}y$[4] have the same operation count.

Now consider parallelizing the product $y \leftarrow Ax$. Suppose we have n processors, and each processor i stores x_i and the i-th row of A. Show that the product Ax can be computed without idle time on any processor but the first.

Can the same be done for the solution of the triangular system $Ax = y$? Show that the straightforward implementation has every processor idle for an $(n-1)/n$ fraction of the computation.

We will now see a number of ways of dealing with this inherently sequential component.

6.9.1 Wavefronts

Above, you saw that solving a lower triangular system of size N can have sequential time complexity of N steps. In practice, things are often not quite that bad. Implicit algorithms such as solving a triangular system are inherently sequential, but the number of steps can be less than is apparent at first.

Exercise 6.20. Take another look at the matrix from a two-dimensional BVP on the unit square, discretized with central differences. Derive the matrix structure if we order the unknowns by diagonals. What can you say about the sizes of the blocks and the structure of the blocks themselves?

Let us take another look at figure 4.1 that describes the difference stencil of a two-dimensional BVP. The corresponding picture for the lower triangular factor is in figure 6.9. This describes the sequentiality of the lower triangular solve process

$$x_k = y_k - \ell_{k,k-1}x_{k-1} - \ell_{k,k-n}x_{k-n}$$

In other words, the value at point k can be found if its neigbours to the left $(k-1)$ and below $(k-n)$ are known. Now we see that, if we know x_1, we can not only find x_2, but also x_{1+n}. In the next step we can determine x_3, x_{n+2}, and x_{2n+1}. Continuing this way, we can solve x by *wavefronts*: the values of x on each wavefront are independent, so they can be solved in parallel in the same sequential step.

4. Obtained by solving the triangular system $Ax = y$, not by inverting A.

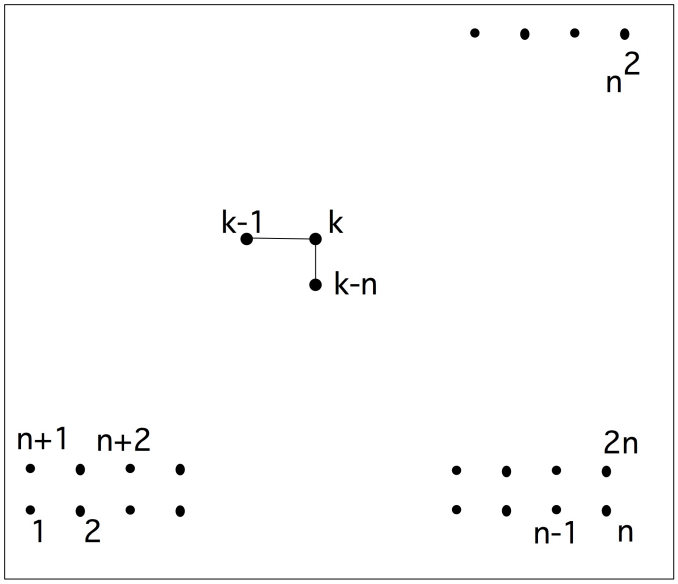


Figure 6.9: The difference stencil of the L factor of the matrix of a two-dimensional BVP

Exercise 6.21. Finish this argument. What is the maximum number of processors we can employ, and what is the number of sequential steps? What is the resulting efficiency?

6.9.2 Recursive doubling

One strategy for dealing with recurrences is *recursive doubling*, which you already saw in exercise 4. Here we will discuss it in a more systematic manner. First, take the matrix from (6.3) and scale it to be of the form

$$\begin{pmatrix} 1 & & \emptyset & \\ b_{21} & 1 & & \\ & \ddots & \ddots & \\ \emptyset & & b_{n,n-1} & 1 \end{pmatrix}$$

which we write as $A = I + B$.

Exercise 6.22. How does solving the system $(I + B)x = y$ help in solving $Ax = y$? What are the operation counts of solving the system in the two different ways?

Now we do something that looks like Gaussian elimination, except that we do not start with the first row, but the second. (What would happen if you did Gaussian elimination or LU decomposition on the matrix $I + B$?) We use

the second row to eliminate b_{32}:

$$\begin{pmatrix} 1 & & & \emptyset \\ & 1 & & \\ & -b_{32} & 1 & \\ & & & \ddots \\ \emptyset & & & 1 \end{pmatrix} \times \begin{pmatrix} 1 & & & \emptyset \\ b_{21} & 1 & & \\ & b_{32} & 1 & \\ & & \ddots & \ddots \\ \emptyset & & b_{n,n-1} & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & \emptyset \\ b_{21} & 1 & & \\ -b_{32}b_{21} & 0 & 1 & \\ \emptyset & & b_{n,n-1} & 1 \end{pmatrix}$$

which we write as $L^{(2)}A = A^{(2)}$. We also compute $L^{(2)}y = y^{(2)}$ so that $A^{(2)}x = y^{(2)}$ has the same solution as $Ax = y$. Solving the transformed system gains us a little: after we compute x_1, x_2 and x_3 can be computed in parallel.

Now we repeat this elimination process by using the fourth row to eliminate b_{54}, the sixth row to eliminate b_{76}, et cetera. The final result is, summarizing all $L^{(i)}$ matrices:

$$\begin{pmatrix} 1 & & & & & & & \emptyset \\ 0 & 1 & & & & & & \\ & -b_{32} & 1 & & & & & \\ & & 0 & 1 & \ddots & & & \\ & & & -b_{54} & 1 & & & \\ & & & & 0 & 1 & & \\ & & & & & -b_{76} & 1 & \\ & & & & & & \ddots & \ddots \end{pmatrix} \times (I+B) = \begin{pmatrix} 1 & & & & & & & \emptyset \\ b_{21} & 1 & & & & & & \\ -b_{32}b_{21} & 0 & 1 & & & & & \\ & & b_{43} & 1 & & & & \\ & & -b_{54}b_{43} & 0 & 1 & & & \\ & & & & b_{65} & 1 & & \\ & & & & -b_{76}b_{65} & 0 & 1 & \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}$$

which we write as $L(I + B) = C$, and solving $(I + B)x = y$ now becomes $Cx = L^{-1}y$.

This final result needs close investigation.

- First of all, computing $y' = L^{-1}y$ is simple. (Work out the details!)
- Solving $Cx = y'$ is still sequential, but it no longer takes n steps: from x_1 we can get x_3, from that we get x_5, et cetera. In other words, there is only a sequential relationship between the odd numbered components of x.
- The even numbered components of x do not depend on each other, but only on the odd components: x_2 follows from x_1, x_4 from x_3, et cetera. Once the odd components have been computed, admittedly sequentially, this step is fully parallel.

We can describe the sequential solving of the odd components by itself:

$$\begin{pmatrix} 1 & & \emptyset \\ c_{21} & 1 & \\ & \ddots & \ddots \\ \emptyset & & c_{n,n-1} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y'_1 \\ y'_3 \\ \vdots \\ y'_n \end{pmatrix}$$

where $c_{i+1i} = -b_{2n+1,2n}b_{2n,2n-1}$. In other words, we have reduced a size n sequential problem to a sequential problem of the size kind and a parallel problem, both of size $n/2$. Now we can repeat this procedure recursively, reducing the original problem to a sequence of parallel operations, each half the size of the former.

6.9.3 Approximating implicit by explicit operations, series expansion

There are various reasons why it is sometimes allowed to replace an implicit operation, which, as you saw above, can be problematic in practice, by a different one that is practically more advantageous.

- Using an explicit method for the heat equation (section 4.3) instead of an implicit one is equally legitimate, as long as we observe step size restrictions on the explicit method.
- Tinkering with the preconditioner (section 5.5.7) in an iterative method is allowed, since it will only affect the speed of convergence, not the solution the method converges to. You already saw one example of this general idea in the *block Jacobi* method; section 6.7.2.1. In the rest of this section, you will see how recurrences, which are implicit operations, can be replaced by explicit operations, giving various computational advantages.

Solving a linear system is a good example of an implicit operation, and since this comes down to solving two triangular systems, let us look at ways of finding a computational alternative to solving a lower triangular system. If U is upper triangular and nonsingular, we let D be the diagonal of U, and we write $U = D(I - B)$ where B is an upper triangular matrix with a zero diagonal, also called a *strictly upper triangular matrix*; we say that $I - B$ is a *unit upper triangular matrix*.

Exercise 6.23. Let $A = LU$ be an LU factorization where L has ones on the diagonal. Show how solving a system $Ax = b$ can be done, involving only the solution of unit upper and lower triangular systems. Show that no divisions are needed during the system solution.

Our operation of interest is now solving the system $(I - B)x = y$. We observe that

$$(I - B)^{-1} = I + B + B^2 + \cdots \tag{6.4}$$

First we observe that $B^n = 0$ where n is the matrix size (check this!), so we can solve $(I - B)x = y$ exactly by

$$x = \sum_{k=0}^{n-1} B^k y.$$

Of course, we want to avoid computing the powers B^k explicitly, so we observe that

$$\sum_{k=0}^{1} B^k y = (I + B)y, \quad \sum_{k=0}^{2} B^k y = (I + B(I + B))y, \quad \sum_{k=0}^{3} B^k y = (I + B(I + B((I + B))))y, \ldots \tag{6.5}$$

The resulting algorithm for evaluating $\sum_{k=0}^{n-1} B^k y$ is called *Horner's rule*, and you see that it avoids computing matrix powers B^k.

Exercise 6.24. Suppose that $I - B$ is bidiagonal. Show that the above calculation takes $n(n + 1)$ operations. What is the operation count for computing $(I - B)x = y$ by triangular solution?

We have now turned an implicit operation into an explicit one, but unfortunately one with a high operation count. In practical circumstances, however, we can truncate the sum of matrix powers.

Exercise 6.25. Let A be the tridiagonal matrix

$$A = \begin{pmatrix} 2 & -1 & & & \emptyset \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & & & -1 \\ \emptyset & & & -1 & 2 \end{pmatrix}$$

of the one-dimensional BVP from section 4.2.2.1.

1. Recall the definition of diagonal dominance in section 5.3.3. Is this matrix diagonally dominant?
2. Show that the pivots in an LU factorization of this matrix (without pivoting) satisfy a recurrence. Hint: show that after n elimination steps ($n \geq 0$) the remaining matrix looks like

$$A^{(n)} = \begin{pmatrix} d_n & -1 & & & \emptyset \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & & & -1 \\ \emptyset & & & -1 & 2 \end{pmatrix}$$

   and show the relation between d_{n+1} and d_n.
3. Show that the sequence $n \mapsto d_n$ is descending, and derive the limit value.
4. Write out the L and U factors in terms of the d_n pivots.
5. Are the L and U factors diagonally dominant?

The above exercise implies (note that we did not actually prove it!) that for matrices from BVPs we find that $B^k \downarrow 0$, in element size and in norm. This means that we can approximate the solution of $(I - B)x = y$ by, for instance, $x = (I+B)y$ or $x = (I+B+B^2)y$. Doing this still has a higher operation count than the direct triangular solution, but it is computationally advantageous in at least two ways:

- The explicit algorithm has a better pipeline behaviour.
- The implicit algorithm has problems in parallel, as you have seen; the explicit algorithm is easily parallelized.

Of course, this approximation may have further implications for the stability of the overall numerical algorithm.

Exercise 6.26. Describe the parallelism aspects of Horner's rule; equation (6.5).

A very similar discussion as the above ensues in the context of computing derivatives through Padé approximations [39]. Here, the systems to be solved are only bi-diagonal, and in the reference cited their solution is replaced by an explicit multiplication by a bi-diagonal.

6.10 Block algorithms on multicore architectures

In section 5.3.7 you saw that certain linear algebra algorithms can be formulated in terms of submatrices. This point of view can be beneficial for the efficient execution of linear algebra operations on shared memory architectures such as current *multicore* processors.

As an example, let us consider the *Cholesky factorization*, which computes $A = LL^t$ for a symmetric positive definite matrix A. Recursively, we can describe the algorithm as follows:

$$\mathrm{Chol}\begin{pmatrix} A_{11} & A_{21}^t \\ A_{21} & A_{22} \end{pmatrix} = LL^t \qquad \text{where} \quad L = \begin{pmatrix} L_{11} & 0 \\ \tilde{A}_{21} & \mathrm{Chol}(A_{22} - \tilde{A}_{21}\tilde{A}_{21}^t) \end{pmatrix}$$

and where $\tilde{A}_{21} = A_{21}L_{11}^{-t}$, $A_{11} = L_{11}L_{11}^t$.

In practice, the block implementation is applied to a partitioning

$$\left(\begin{array}{c|cc} & \multicolumn{2}{c}{\text{finished}} \\ \hline & A_{kk} & A_{k,>k} \\ \hline & A_{>k,k} & A_{>k,>k} \end{array}\right)$$

where k is the index of the current block row, and the factorization is finished for all indices $< k$. The factorization is written as follows, using Blas names for the operations:

for $k = 1, \mathrm{nblocks}$:
- `Chol`: factor $A_{kk} = L_kL_k^t$
- `Trsm`: solve $\tilde{A}_{>k,k} = A_{>k,k}L_k^{-t}$
- `Gemm`: multiply $\tilde{A}_{>k,k}\tilde{A}_{>k,k}^t$
- `Syrk`: symmmetric rank-k update $A_{>k,>k} \leftarrow A_{>k,>k} - \tilde{A}_{>k,k}\tilde{A}_{>k,k}^t$

The key to parallel performance is to partition the indices $> k$ and write the algorithm in terms of these blocks:

$$\left(\begin{array}{c|cc} & \multicolumn{2}{c}{\text{finished}} \\ \hline & A_{kk} & A_{k,k+1} \quad A_{k,k+2}\cdots \\ \hline & A_{k+1,k} & A_{k+1,k+1} \quad A_{k+1,k+2}\cdots \\ & A_{k+2,k} & A_{k+2,k+2} \\ & \vdots & \vdots \end{array}\right)$$

The algorithm now gets an extra inner loop:

for $k = 1, \mathrm{nblocks}$:
- `Chol`: factor $A_{kk} = L_kL_k^t$
- for $\ell > k$:
  - `Trsm`: solve $\tilde{A}_{\ell,k} = A_{\ell,k}L_k^{-t}$
- for $\ell_1, \ell_2 > k$:
  - `Gemm`: multiply $\tilde{A}_{\ell_1,k}\tilde{A}_{\ell_2,k}^t$
- for $\ell_1, \ell_2 > k$, $\ell_1 \leq \ell_2$:
  - `Syrk`: symmmetric rank-k update $A_{\ell_1,\ell_2} \leftarrow A_{\ell_1,\ell_2} - \tilde{A}_{\ell_1,k}\tilde{A}_{\ell_2,k}^t$

Now it is clear that the algorithm has a good deal of parallelism: the iterations in every ℓ-loop can be processed independently. However, these loops get shorter in every iteration of the outer k-loop, so it is not immediate how many processors we can accomodate. Moreover, it is not necessary to preserve the order of operations of the algorithm above. For instance, after

$$L_1 L_1^t = A_{11}, \quad A_{21} \leftarrow A_{21} L_1^{-t}, \quad A_{22} \leftarrow A_{22} - A_{21} A_{21}^t$$

the factorization $L_2 L_2^t = A_{22}$ can start, even if the rest of the $k = 1$ iteration is still unfinished. Instead of looking at the algorithm, it is a better idea to construct a Directed Acyclic Graph (DAG) (see section A.6 for a brief tutorial on graphs) of the tasks of all inner iterations. Figure 6.10 shows the DAG of all tasks of matrix of 4×4 blocks. This graph is constructed by simulating the Cholesky algorithm above, making a vertex for every task, adding an edge (i, j) if task j uses the output of task i.

Exercise 6.27. What is the diameter of this graph? Identify the tasks that lie on the path that determines the diameter. What is the meaning of these tasks in the context of the algorithm? This path is called the 'critical path'. Its length determines the execution time of the computation in parallel, even if an infinite number of processors is available.

Exercise 6.28. If there are T tasks – all taking unit time to execute – and we have p processors, what is the theoretical minimum time to execute the algorithm? Now amend this formula to take into account the critical path; call its length C.

In the execution of the tasks a DAG, several observations can be made.

- If more than one update is made to a block, it is probably advantageous to have these updates be computed by the same process. This simplifies maintaining *cache coherence*.
- If data is used and later modified, the use must be finished before the modification can start. This can even be true if the two actions are on different processors, since the memory subsystem typically maintains cache coherence, so the modifications can affect the process that is reading the data. This case can be remedied by having a copy of the data in main memory, giving a reading process data that is reserved (see section 1.3.1).

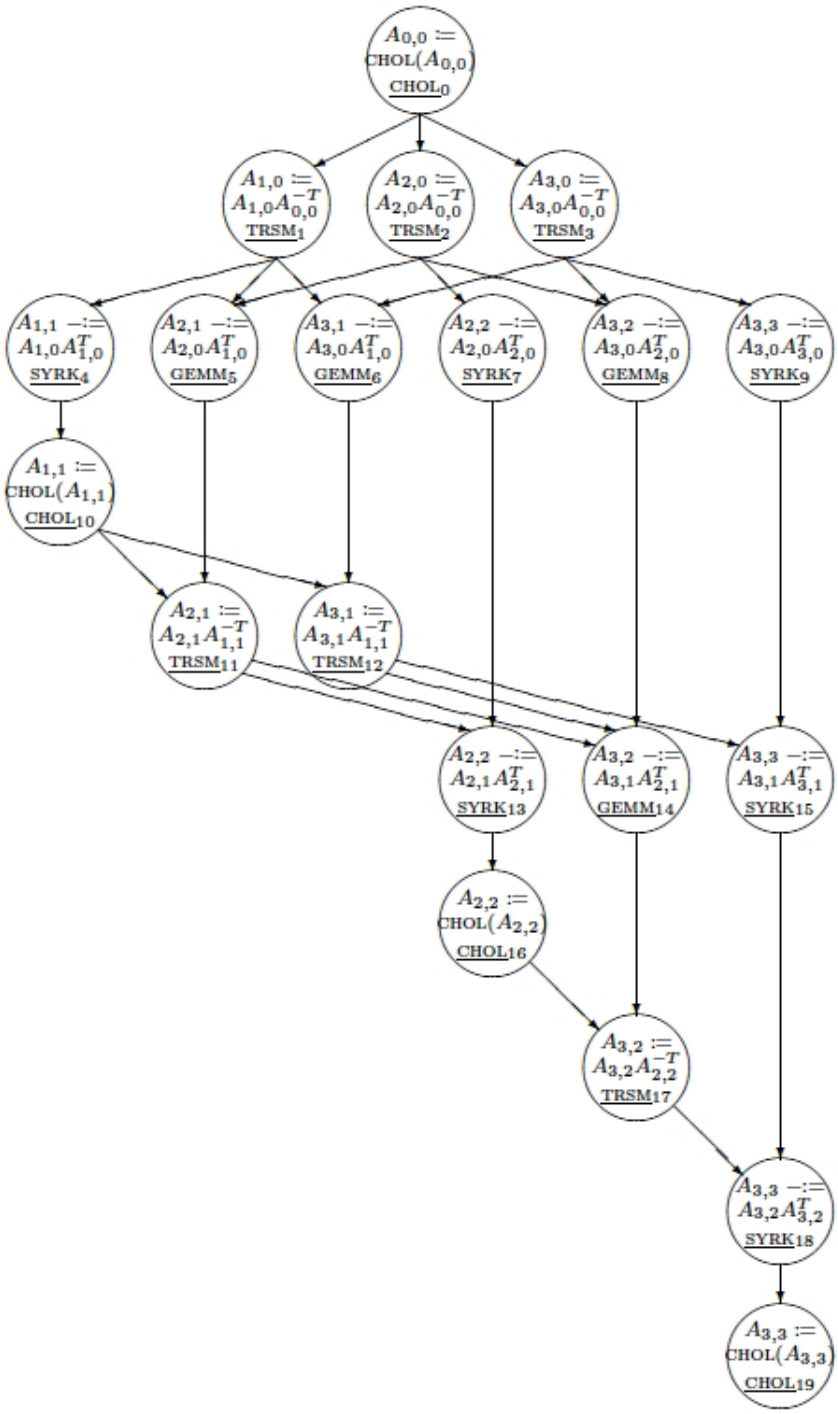


Figure 6.10: Graph of task dependencies in a 4×4 Cholesky factorization

Chapter 7

Molecular dynamics

Molecular dynamics is a technique for simulating the atom-by-atom behavior of molecules and deriving macroscopic properties from these atomistic motions. It has application to biological molecules such as proteins and nucleic acids, as well as natural and synthetic molecules in materials science and nanotechnology. Molecular dynamics falls in the category of particle methods, which includes N-body problems in celestial mechanics and astrophysics, and many of the ideas presented here will carry over to these other fields. In addition, there are special cases of molecular dynamics including ab initio molecular dynamics where electrons are treated quantum mechanically and thus chemical reactions can be modeled. We will not treat these special cases, but will instead concentrate on *classical* molecular dynamics.

The idea behind molecular dynamics is very simple: a set of particles interact according to Newton's law of motion, $F = ma$. Given the initial particle positions and velocities, the particle masses and other parameters, as well as a model of the forces that act between particles, Newton's law of motion can be integrated numerically to give a trajectory for each of the particles for all future (and past) time. Commonly, the particles reside in a computational box with periodic boundary conditions.

A molecular dynamics time step is thus composed of two parts:

1: compute forces on all particles
2: update positions (integration).

The computation of the forces is the expensive part. State-of-the-art molecular dynamics simulations are performed on parallel computers because the force computation is costly and a vast number of time steps are required for reasonable simulation lengths. In many cases, molecular dynamics is applied to simulations on molecules with a very large number of atoms as well, e.g., up to a million for biological molecules and long time scales, and up to billions for other molecules and shorter time scales.

Numerical integration techniques are also of interest in molecular dynamics. For simulations that take a large number of time steps and for which the preservation of quantities such as energy is more important than order of accuracy, the solvers that must be used are different than the traditional ODE solvers presented in Chapter 4.

In the following, we will introduce force fields used for biomolecular simulations and discuss fast methods for computing these forces. Then we devote sections to the parallelization of molecular dynamics for short-range forces

and the parallelization of the 3-D FFT used in fast computations of long-range forces. We end with a section introducing the class of integration techniques that are suitable for molecular dynamics simulations. Our treatment of the subject of molecular dynamics in this chapter is meant to be introductory and practical; for more information, the text [34] is recommended.

7.1 Force Computation

7.1.1 Force Fields

In classical molecular dynamics, the model of potential energy and of the forces that act between atoms is called a *force field*. The force field is a tractable but approximate model of quantum mechanical effects which are computationally too expensive to determine for large molecules. Different force fields are used for different types of molecules, as well as for the same molecule by different researchers, and none are ideal.

In biochemical systems, commonly-used force fields model the potential energy function as the sum of bonded, van der Waals, and electrostatic (Coulomb) energy:

$$E = E_{\text{bonded}} + E_{\text{Coul}} + E_{\text{vdW}}.$$

The potential is a function of the positions of all the atoms in the simulation. The force on an atom is the negative gradient of this potential at the position of the atom.

The bonded energy is due to covalent bonds in a molecule,

$$E_{\text{bonded}} = \sum_{\text{bonds}} k_i(r_i - r_{i,0})^2 + \sum_{\text{angles}} k_i(\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} V_n(1 + \cos(n\omega - \gamma))$$

where the three terms are, respectively, sums over all covalent bonds, sums over all angles formed by two bonds, and sums over all dihedral angles formed by three bonds. The fixed parameters k_i, $r_{i,0}$, etc. depend on the types of atoms involved, and may differ for different force fields. Additional terms or terms with different functional forms are also commonly used.

The remaining two terms for the potential energy E are collectively called the nonbonded terms. Computationally, they form the bulk of the force calculation. The electrostatic energy is due to atomic charges and is modeled by the familiar

$$E_{\text{Coul}} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

where the sum is over all pairs of atoms, q_i and q_j are the charges on atoms i and j, and r_{ij} is the distance between atoms i and j. Finally, the van der Waals energy approximates the remaining attractive and repulsive effects, and is commonly modeled by the Lennard-Jones function

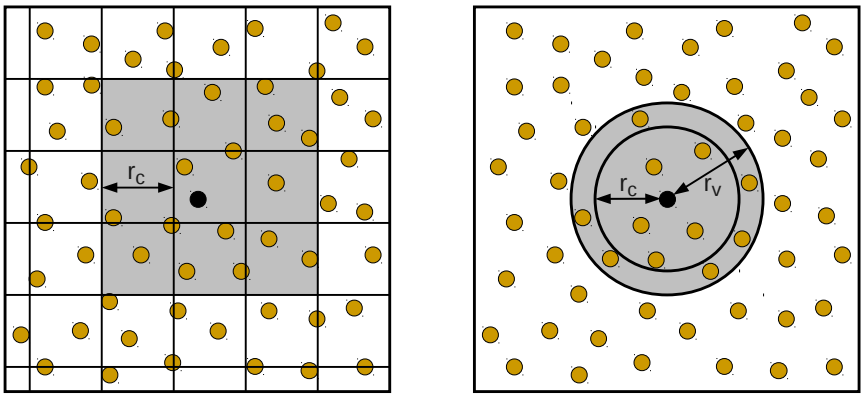
$$E_{\text{vdW}} = \sum_i \sum_{j>i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

where ϵ_{ij} and σ_{ij} are force field parameters depending on atom types. At short distances, the repulsive (r^{12}) term is in effect, while at long distances, the dispersive (attractive, $-r^6$) term is in effect.

Parallelization of the molecular dynamics force calculation depends on parallelization each of these individual types of force calculations. The bonded forces are local computations in the sense that for a given atom, only nearby atom positions and data are needed. The van der Waals forces are also local and are termed short-range because they are negligible for large atom separations. The electrostatic forces are long-range, and various techniques have been developed to speed up these calculations. In the next two subsections, we separately discuss the computation of short-range and long-range nonbonded forces.

7.1.2 Computing Short-Range Nonbonded Forces

The computation of short-range nonbonded forces for a particle can be truncated beyond a cutoff radius, r_c, of that particle. The naive approach to perform this computation for a particle i is by examining all other particles and computing their distance to particle i. For n particles, the complexity of this approach is $O(n^2)$, which is equivalent to computing forces between all pairs of particles. There are two data structures, *cell lists* and *Verlet neighbor lists*, that can be used independently for speeding up this calculation, as well as an approach that combines the two.



(a) Cell list method. (b) Verlet neighbor list method.

Figure 7.1: Computing nonbonded forces within a cutoff, r_c. To compute forces involving the highlighted particle, only particles in the shaded regions are considered.

Cell Lists

The idea of cell lists appears often in problems where a set of points that are nearby a given point is sought. Referring to Fig. 7.1(a), where we illustrate the idea with a 2-D example, a grid is laid over the set of particles. If the grid spacing is no less than r_c, then to compute the forces on particle i, only the particles in the cell containing i and the 8 adjacent cells need to be considered. One sweep through all the particles is used to construct a list of particles for each cell. These cell lists are used to compute the forces for all particles. At the next time step, since the particles

have moved, the cell lists must be regenerated or updated. The complexity of this approach is $O(n)$ for computing the data structure and $O(n \times n_c)$ for the force computation, where n_c is the average number of particles in 9 cells (27 cells in 3-D). The storage required for the cell list data structure is $O(n)$.

Verlet Neighbor Lists

The cell list structure is somewhat inefficient because, for each particle i, n_c particles are considered, but this is much more than the number of particles within the cutoff r_c. A Verlet neighbor list is a list of particles within the cutoff for a particle i. Each particle has its own list, and thus the storage required is $O(n \times n_v)$ where n_v is the average number of particles within the cutoff. Once these lists are constructed, computing the forces is then very fast, requiring the minimal complexity $O(n \times n_v)$. Constructing the list is more expensive, requiring examining all the particles for each particle, i.e., no less than the original complexity of $O(n^2)$. The advantage, however, is that the neighbor lists can be reused for many time steps if an expanded cutoff, r_v is used. Referring to a 2-D example in Fig. 7.1(b), the neighbor list can be reused as long as no particle from outside the two circles moves inside the inner circle. If the maximum speed of the particles can be estimated or bounded, then one can determine a number of time steps for which it is safe to reuse the neighbor lists. (Alternatively, it may be possible to signal when any particle crosses to a position within the cutoff.) Technically, the Verlet neighbor list is the list of particles within the expanded cutoff, r_v.

Using Cell and Neighbor Lists Together

The hybrid approach is simply to use Verlet neighbor lists but to use cell lists to construct the neighbor lists. This reduces the high cost when neighbor lists need to be regenerated. This hybrid approach is very effective and is often the approach used in state-of-the-art molecular dynamics software.

Both cell lists and Verlet neighbor lists can be modified to exploit the fact that the force f_{ij} on particle i due to particle j is equal to $-f_{ji}$ (Newton's third law) and only needs to be computed once. For example, for cell lists, only 4 of the 8 cells (in 2-D) need to be considered.

7.1.3 Computing Long-Range Forces

Electrostatic forces are challenging to compute because they are long-range: each particle feels a non-negligible electrostatic force from all other particles in the simulation. An approximation that is sometimes used is to truncate the force calculation for a particle after a certain cutoff radius (as is done for short-range van der Waals forces). This generally produces unacceptable artifacts in the results, however.

There are several more accurate methods for speeding up the computation of electrostatic forces, avoiding the $O(n^2)$ sum over all pairs of n particles. We briefly outline some of these methods here.

Hierarchical N-body Methods

Hierarchical N-body methods, including the Barnes-Hut method and the fast multipole method, are popular for astrophysical particle simulations, but are typically too costly for the accuracy required in biomolecular simulations.

In the Barnes-Hut method, space is recursively divided into 8 equal cells (in 3-D) until each cell contains zero or one particles. Forces between nearby particles are computed individually, as normal, but for distant particles, forces are computed between one particle and a set of distant particles within a cell. An accuracy measure is used to determine if the force can be computed using a distant cell or must be computed by individually considering its children cells. The Barnes-Hut method has complexity $O(n \log n)$. The fast multipole method has complexity $O(n)$; this method calculates the potential and does not calculate forces directly.

Particle-Mesh Methods

In particle-mesh methods, we exploit the Poisson equation

$$\nabla^2 \phi = -\frac{1}{\epsilon}\rho$$

which relates the potential ϕ to the charge density ρ, where $1/\epsilon$ is a constant of proportionality. To utilize this equation, we discretize space using a mesh, assign charges to the mesh points, solve Poisson's equation on the mesh to arrive at the potential on the mesh. The force is the negative gradient of the potential (for conservative forces such as electrostatic forces). A number of techniques have been developed for distributing point charges in space to a set of mesh points and also for numerically interpolating the force on the point charges due to the potentials at the mesh points. Many fast methods are available for solving the Poisson equation, including multigrid methods and fast Fourier transforms. With respect to terminology, particle-mesh methods are in contrast to the naive *particle-particle* method where forces are computed between all pairs of particles.

It turns out that particle-mesh methods are not very accurate, and a more accurate alternative is to split each force into a short-range, rapidly-varying part and a long-range, slowly-varying part:

$$f_{ij} = f_{ij}^{sr} + f_{ij}^{lr}.$$

One way to accomplish this easily is to weigh f by a function $h(r)$, which emphasizes the short-range part (small r) and by $1 - h(r)$ which emphasizes the long-range part (large r). The short-range part is computed by computing the interaction of all pairs of particles within a cutoff (a particle-particle method) and the long-range part is computed using the particle-mesh method. The resulting method, called particle-particle-particle-mesh (PPPM, or P^3M) is due to Hockney and Eastwood, in a series of papers beginning in 1973.

Ewald Method

The Ewald method is the most popular of the methods described so far for electrostatic forces in biomolecular simulations and was developed for the case of periodic boundary conditions. The structure of the method is similar to PPPM in that the force is split between short-range and long-range parts. Again, the short-range part is computed using particle-particle methods, and the long-range part is computed using Fourier transforms. Variants of the Ewald method are very similar to PPPM in that the long-range part uses a mesh, and fast Fourier transforms are used to solve the Poisson equation on the mesh. For additional details, see, for example [34]. In Section 7.3, we describe the parallelization of the 3-D FFT to solve the 3-D Poisson equation.

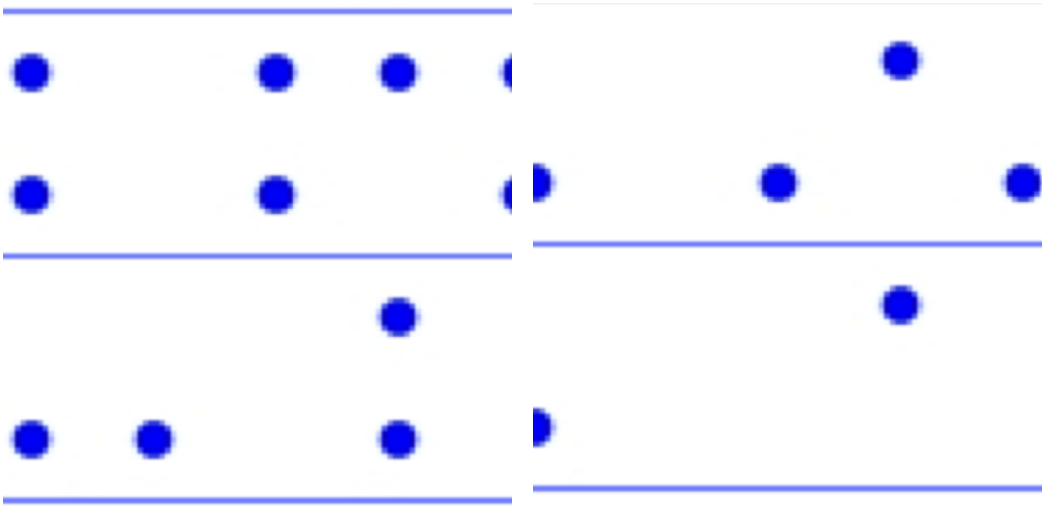
Victor Eijkhout

7.2 Parallel Decompositions

We now discuss the parallel computation of forces. Plimpton [70] created a very useful categorization of molecular dynamics parallelization methods, identifying *atom*, *force*, and *spatial* decomposition methods. Here, we closely follow his description of these methods. We also add a fourth category which has come to be recognized as differing from the earlier categories, called *neutral territory* methods, a name coined by Shaw [74]. Neutral territory methods are currently used by many state-of-the-art molecular dynamics codes. Spatial decompositions and neutral territory methods are particularly advantageous for parallelizing cutoff-based calculations.

7.2.1 Atom Decompositions

In an atom decomposition, each particle is assigned to one processor, and that processor is responsible for computing the particle's forces and updating its position for the entire simulation. For the computation to be roughly balanced, each processor is assigned approximately the same number of particles (a random distribution works well). An important point of atom decompositions is that each processor generally needs to communicate with all other processors to share updated particle positions.



(a) Force matrix. (b) Force matrix, redundancies removed.

Figure 7.2: Atom decomposition, showing a force matrix of 16 particles distributed among 8 processors. A dot represents a nonzero entry in the force matrix. On the left, the matrix is symmetric; on the right, only one element of a pair of skew-symmetric elements is computed, to take advantage of Newton's third law.

An atom decomposition is illustrated by the *force matrix* in Fig. 7.2(a). For n particles, the force matrix is an n-by-n matrix; the rows and columns are numbered by particle indices. A nonzero entry f_{ij} in the matrix denotes a nonzero force on particle i due to particle j which must be computed. This force may be a nonbonded and/or a bonded force. When cutoffs are used, the matrix is sparse, as in this example. The matrix is dense if forces are computed

between all pairs of particles. The matrix is skew-symmetric because of Newton's third law, $f_{ij} = -f_{ji}$. The lines in Fig. 7.2(a) show how the particles are partitioned. In the figure, 16 particles are partitioned among 8 processors.

Algorithm 1 shows one time step from the point of view of one processor. At the beginning of the time step, each processor holds the positions of particles assigned to it.

Algorithm 1 Atom decomposition time step

1: send/receive particle positions to/from all other processors
2: (if nonbonded cutoffs are used) determine which nonbonded forces need to be computed
3: compute forces for particles assigned to this processor
4: update positions (integration) for particles assigned to this processor

An optimization is to halve the amount of computation, which is possible because the force matrix is skew-symmetric. To do this, we choose exactly one of f_{ij} or f_{ji} for all skew-symmetric pairs such that each processor is responsible for computing approximately the same number of forces. Choosing the upper or lower triangular part of the force matrix is a bad choice because the computational load is unbalanced. A better choice is to compute f_{ij} if $i + j$ is even in the upper triangle, or if $i + j$ is odd in the lower triangle, as shown in Fig. 7.2(b). There are many other options.

When taking advantage of skew-symmetry in the force matrix, all the forces on a particle owned by a processor are no longer computed by that processor. For example, in Fig. 7.2(b), the forces on particle 1 are no longer computed only by the first processor. To complete the force calculation, processors must communicate to send forces that are needed by other processors and receive forces that are computed by other processors. The above algorithm must now be modified by adding a communication step (step 4) as shown in Algorithm 2.

Algorithm 2 Atom decomposition time step, without redundant calculations

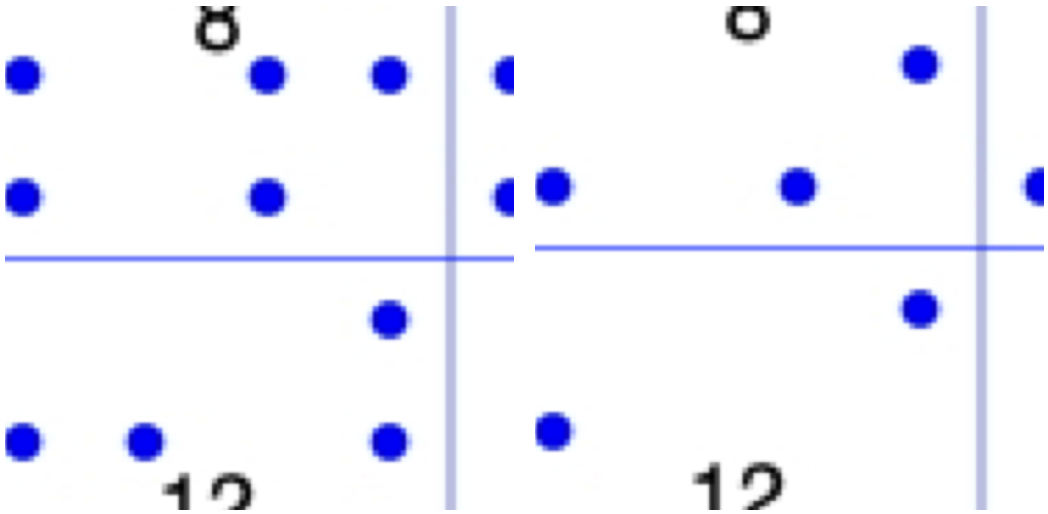
1: send/receive particle positions to/from all other processors
2: (if nonbonded cutoffs are used) determine which nonbonded forces need to be computed
3: compute *partial* forces for particles assigned to this processor
4: send particle forces needed by other processors and receive particle forces needed by this processor
5: update positions (integration) for particles assigned to this processor

This algorithm is advantageous if the extra communication is outweighed by the savings in computation. Note that the amount of communication doubles in general.

7.2.2 Force Decompositions

In a force decomposition, the forces are distributed among the processors for computation. A straightforward way to do this is to partition the force matrix into blocks and to assign each block to a processor. Fig. 7.3(a) illustrates this for the case of 16 particles and 16 processors. Particles also need to be assigned to processors (as in atom decompositions) for the purpose of having processors assigned to update particle positions. In the example of the

Figure, processor i is assigned to update the positions of particle i; in practical problems, a processor would be assigned to update the positions of many particles. Note that, again, we first consider the case of a skew-symmetric force matrix.



(a) Force matrix. (b) Force matrix, redundancies removed.

Figure 7.3: Force decomposition, showing a force matrix of 16 particles and forces partitioned among 16 processors.

We now examine the communication required in a time step for a force decomposition. Consider processor 3, which computes partial forces for particles 0, 1, 2, 3, and needs positions from particles 0, 1, 2, 3, and also 12, 13, 14, 15. Thus processor 3 needs to perform communication with processors 0, 1, 2, 3, and processors 12, 13, 14, 15. After forces have been computed by all processors, processor 3 needs to collect forces on particle 3 computed by other processors. Thus processor 2 needs to perform communication again with processors 0, 1, 2, 3.

Algorithm 3 shows what is performed in one time step, from the point-of-view of one processor. At the beginning of the time step, each processor holds the positions of all the particles assigned to it.

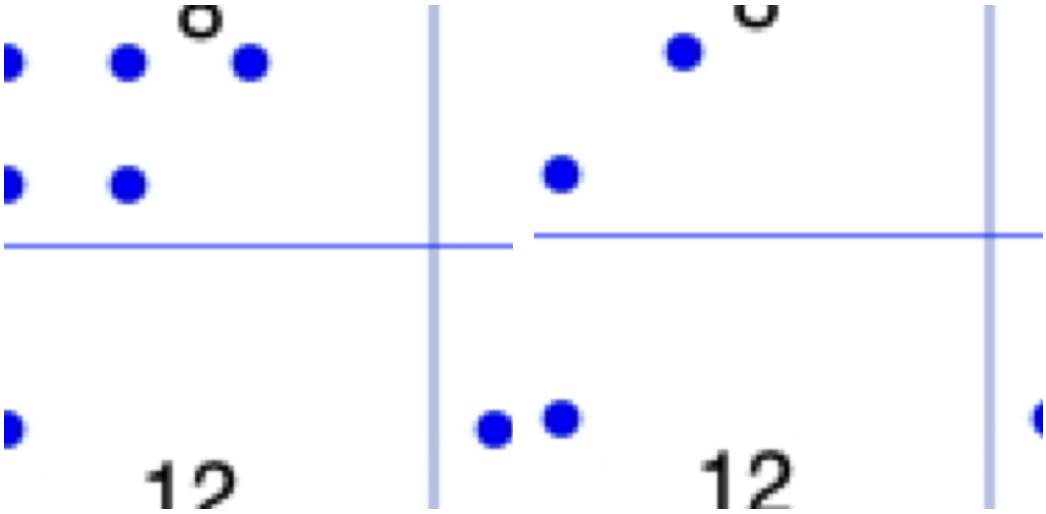
In general, if there are p processors (and p is square, for simplicity), then the the force matrix is partitioned into $\sqrt{p}$ by $\sqrt{p}$ blocks. The force decomposition just described requires a processor to communicate in three steps, with $\sqrt{p}$ processors in each step. This is much more efficient than atom decompositions which require communications among all p processors.

We can also exploit Newton's third law in force decompositions. Like for atom decompositions, we first choose a modified force matrix where only one of f_{ij} and f_{ji} is computed. The forces on particle i are computed by a row of processors and now also by a column of processors. Thus an extra step of communication is needed by each processor to collect forces from a column of processors for particles assigned to it. Whereas there were three communication steps, there are now four communication steps when Newton's third law is exploited (the communication is not doubled in this case as in atom decompositions).

Algorithm 3 Force decomposition time step

1: send positions of my assigned particles which are needed by other processors; receive *row* particle positions needed by my processor (this communication is between processors in the same processor row, e.g., processor 3 communicates with processors 0, 1, 2, 3)
2: receive *column* particle positions needed by my processor (this communication is generally with processors in another processor row, e.g., processor 3 communicates with processors 12, 13, 14, 15)
3: (if nonbonded cutoffs are used) determine which nonbonded forces need to be computed
4: compute forces for my assigned particles
5: send forces needed by other processors; receive forces needed for my assigned particles (this communication is between processors in the same processor row, e.g., processor 3 communicates with processors 0, 1, 2, 3)
6: update positions (integration) for my assigned particles

A modification to the force decomposition saves some communication. In Fig. 7.4, the columns are reordered using a *block-cyclic* ordering. Consider again processor 3, which computes partial forces for particles 0, 1, 2, 3. It needs positions from particles 0, 1, 2, 3, as before, but now also with processors 3, 7, 11, 15. The latter are processors in the same processor column as processor 3. Thus all communications are within the same processor row or processor column, which may be advantageous on mesh-based network architectures. The modified method is shown as Algorithm 4.



(a) Force matrix. (b) Force matrix, redundancies removed.

Figure 7.4: Force decomposition, with permuted columns in the force matrix. Note that columns 3, 7, 11, 15 are now in the block column corresponding to processors 3, 7, 11, 15 (the same indices), etc.

Algorithm 4 Force decomposition time step, with permuted columns of force matrix

1: send positions of my assigned particles which are needed by other processors; receive *row* particle positions needed by my processor (this communication is between processors in the same processor row, e.g., processor 3 communicates with processors 0, 1, 2, 3)
2: receive *column* particle positions needed by my processor (this communication is generally with processors the same processor column, e.g., processor 3 communicates with processors 3, 7, 11, 15)
3: (if nonbonded cutoffs are used) determine which nonbonded forces need to be computed
4: compute forces for my assigned particles
5: send forces needed by other processors; receive forces needed for my assigned particles (this communication is between processors in the same processor row, e.g., processor 3 communicates with processors 0, 1, 2, 3)
6: update positions (integration) for my assigned particles

7.2.3 Spatial Decompositions

In a spatial decomposition, space is decomposed into cells. Each cell is assigned to a processor which is responsible for computing the forces on particles that lie inside the cell. Fig. 7.5(a) illustrates a spatial decomposition into 64 cells for the case of a 2-D simulation. (This is a decomposition of space and is not to be confused with a force matrix.) Typically, the number of cells is chosen to be equal to the number of processors. Since particles move during the simulation, the assignment of particles to cells changes as well. This is in contrast to atom and force decompositions.

Fig. 7.5(b) shows one cell (center square) and the region of space (shaded) that contains particles that are potentially within the cutoff radius, r_c, with particles in the given cell. The shaded region is often called the *import region*, since the given cell must import positions of particles lying in this region to perform its force calculation. Note that not all particles in the given cell must interact with all particles in the import region, especially if the import region is large compared to the cutoff radius.

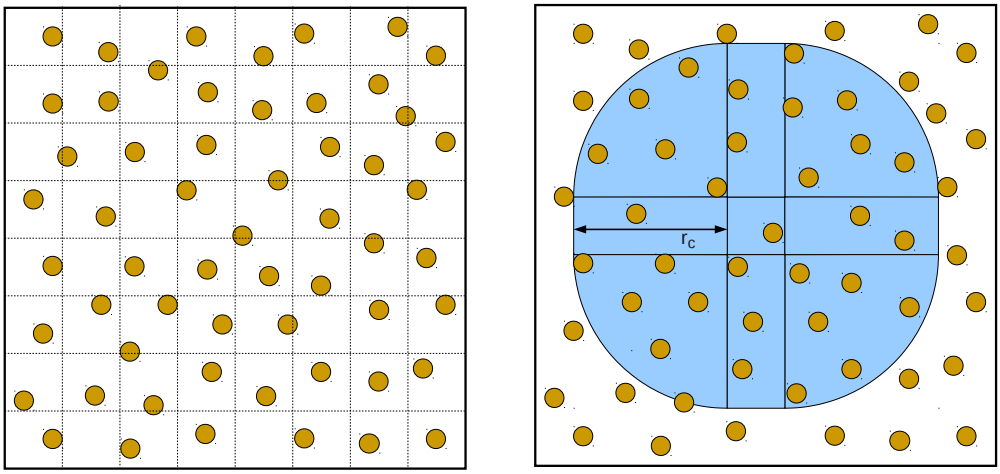
Algorithm 5 shows what each processor performs in one time step. We assume that at the beginning of the time step, each processor holds the positions of the particles in its cell.

Algorithm 5 Spatial decomposition time step

1: send positions needed by other processors for particles in their import regions; receive positions for particles in my import region
2: compute forces for my assigned particles
3: update positions (integration) for my assigned particles

To exploit Newton's third law, the shape of the import region can be halved. Now each processor only computes a partial force on particles in its cell, and needs to receive forces from other processors to compute the total force on these particles. Thus an extra step of communication is involved. We leave it as an exercise to the reader to work out the details of the modified import region and the pseudocode for this case.

In the implementation of a spatial decomposition method, each cell is associated with a list of particles in its import



(a) Decomposition into 64 cells. (b) Import region for one cell.

Figure 7.5: Spatial decomposition, showing particles in a 2-D computational box, (a) partitioned into 64 cells, (b) import region for one cell.

region, similar to a Verlet neighbor list. Like a Verlet neighbor list, it is not necessary to update this list at every time step, if the import region is expanded slightly. This allows the import region list to be reused for several time steps, corresponding to the amount of time it takes a particle to traverse the width of the expanded region. This is exactly analogous to Verlet neighbor lists.

In summary, the main advantage of spatial decomposition methods is that they only require communication between processors corresponding to nearby particles. A disadvantage of spatial decomposition methods is that, for very large numbers of processors, the import region is large compared to the number of particles contained inside each cell.

7.2.4 Neutral Territory Methods

Our description of neutral territory methods follows closely that of Shaw [74]. A neutral territory method can be viewed as combining aspects of spatial decompositions and force decompositions. To parallelize the integration step, particles are assigned to processors according to a partitioning of space. To parallelize the force computation, each processor computes the forces between two sets of particles, but these particles may be unrelated to the particles that have been assigned to the processor for integration. As a result of this additional flexibility, neutral territory methods may require much less communication than spatial decomposition methods.

An example of a neutral territory method is shown in Fig. 7.6 for the case of a 2-D simulation. In the method shown in the Figure, the given processor is assigned the computation of forces between particles lying in the horizontal bar with particles lying in the vertical bar. These two regions thus form the import region for this method. By comparing to Fig. 7.6(b), the import region for this neutral territory method is much smaller than that for the corresponding

spatial decomposition method. The advantage is greater when the size of the cells corresponding to each processor is small compared to the cutoff radius.

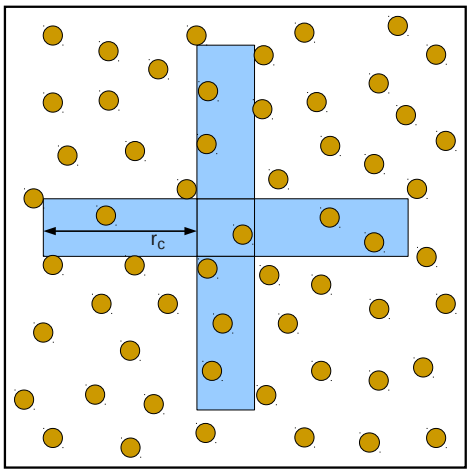


Figure 7.6: Neutral territory method, showing particles in a 2-D computational box and the import region (shaded) for one cell (center square). This Figure can be compared directly to the spatial decomposition case of Fig. 7.5(b). See Shaw [74] for additional details.

After the forces are computed, the given processor sends the forces it has computed to the processors that need these forces for integration. We thus have Algorithm 6.

Algorithm 6 Neutral territory method time step

1: send and receive particle positions corresponding to import regions
2: compute forces assigned to this processor
3: send and receive forces required for integration
4: update positions (integration) for particles assigned to this processor

Like other methods, the import region of the neutral territory method can be modified to take advantage of Newton's third law. We refer to Shaw [74] for additional details and for illustrations of neutral territory methods in 3-D simulations.

7.3 Parallel Fast Fourier Transform

A common component of many methods for computing long-range forces is the 3-D FFT for solving the Poisson equation on a 3-D mesh. The Fourier transform diagonalizes the Poisson operator (called the Laplacian) and one forward and one inverse FFT transformation are required in the solution. Consider the discrete Laplacian operator

L (with periodic boundary conditions) and the solution of ϕ in $-L\phi = \rho$. Let F denote the Fourier transform. The original problem is equivalent to

$$\begin{aligned} -(FLF^{-1})F\phi &= F\rho \\ \phi &= -F^{-1}(FLF^{-1})^{-1}F\rho. \end{aligned}$$

The matrix FLF^{-1} is diagonal. The forward Fourier transform F is applied to ρ, then the Fourier-space components are scaled by the inverse of the diagonal matrix, and finally, the inverse Fourier transform F^{-1} is applied to obtain the solution ϕ.

For realistic protein sizes, a mesh spacing of approximately 1 Ångstrom is typically used, leading to a 3-D mesh that is quite small by many standards: $64 \times 64 \times 64$, or $128 \times 128 \times 128$. Parallel computation would often not be applied to a problem of this size, but parallel computation must be used because the data ρ is already distributed among the parallel processors (assuming a spatial decomposition is used).

A 3-D FFT is computed by computing 1-D FFTs in sequence along each of the three dimensions. For the $64 \times 64 \times 64$ mesh size, this is 4096 1-D FFTs of dimension 64. The parallel FFT calculation is typically bound by communication. The best parallelization of the FFT depends on the size of the transforms and the architecture of the computer network. Below, we first describe some concepts for parallel 1-D FFTs and then describe some concepts for parallel 3-D FFTs. For current software and research dedicated to the parallelization and efficient computation (using SIMD operations) of large 1-D transforms, we refer to the SPIRAL and FFTW packages. These packages use autotuning to generate FFT codes that are efficient for the user's computer architecture.

7.3.1 Parallel 1-D FFT

1-D FFT without Transpose

Fig. 7.7 shows the data dependencies (data flow diagram) between the inputs (left) and outputs (right) for the 16-point radix-2 decimation-in-frequency FFT algorithm. (Not shown is a bit-reversal permutation that may be necessary in the computation.) The Figure also shows a partitioning of the computation among four processors. In this parallelization, the initial data is not moved among processors, but communication occurs during the computation. In the example shown in the Figure, communication occurs in the first two FFT stages; the final two stages do not involve communication. When communication does occur, every processor communicates with exactly one other processor.

1-D FFT with Transpose

Use of transposes is common to parallelize FFT computations. Fig. 7.8(a) shows the same data flow diagram as in Fig. 7.7, but horizontal lines have been removed and additional index labels have been added for clarity. As before, the first two FFT stages are performed without communication. The data is then transposed among the processors. With this transposed data layout, the last two FFT stages can be performed without communication. The final data is not in the original order; an additional transpose may be needed, or the data may be used in this transposed order. Fig. 7.8(b) shows how the indices are partitioned among four processors before and after the transpose. From these

two Figures, notice that the first two stages have data dependencies that only involve indices in the same partition. The same is true for the second two stages for the partitioning after the transpose. Observe also that the structure of the computations before and after the transpose are identical.

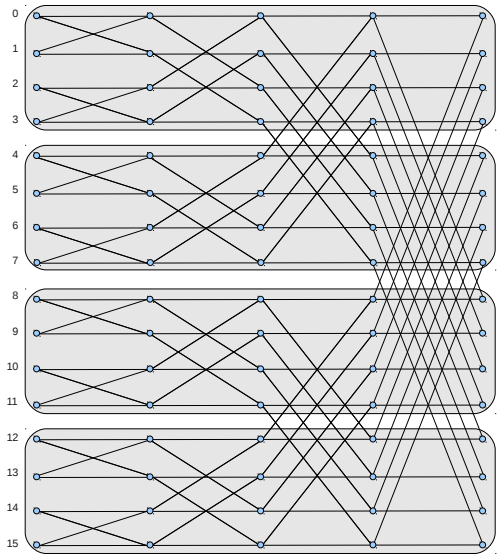


Figure 7.7: Data flow diagram for 1-D FFT for 16 points. The shaded regions show a decomposition for 4 processors (one processor per region). In this parallelization, the first two FFT stages have no communication; the last two FFT stages do have communication.

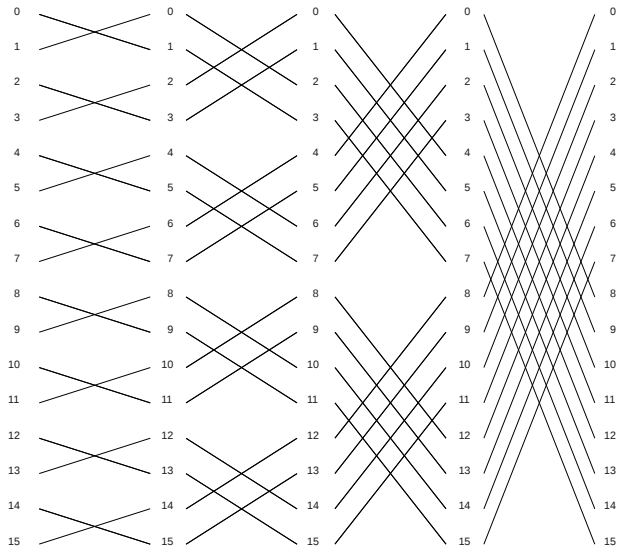
7.3.2 Parallel 3-D FFT

3-D FFT with Block Decomposition

Fig. 7.9(a) shows a block decomposition of the FFT input data when a spatial decomposition is used for a mesh of size $8 \times 8 \times 8$ distributed across 64 processors arranged in a $4 \times 4 \times 4$ topology. The parallel 1-D FFT algorithms can be applied in each of the dimensions. For the example shown in the Figure, each 1-D FFT computation involves 4 processors. Each processor performs multiple 1-D FFTs simultaneously (four in this example). Within each processor, data is ordered contiguously if traversing one of the dimensions, and thus data access is strided for computation in the other two dimensions. Strided data access can be slow, and thus it may be worthwhile to reorder the data within each processor when computing the FFT for each of the dimensions.

3-D FFT with Slab Decomposition

The slab decomposition is shown in Fig. 7.9(b) for the case of 4 processors. Each processor holds one or more planes of the input data. This decomposition is used if the input data is already distributed in slabs, or if it can be



(a) Data flow diagram (shown without horizontal lines for clarity) for 1-D FFT for 16 points.

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

0	4	8	12
1	5	9	13
2	6	10	14
3	7	11	15

(b) Partitioning of the indices before (left) and after (right) the transpose.

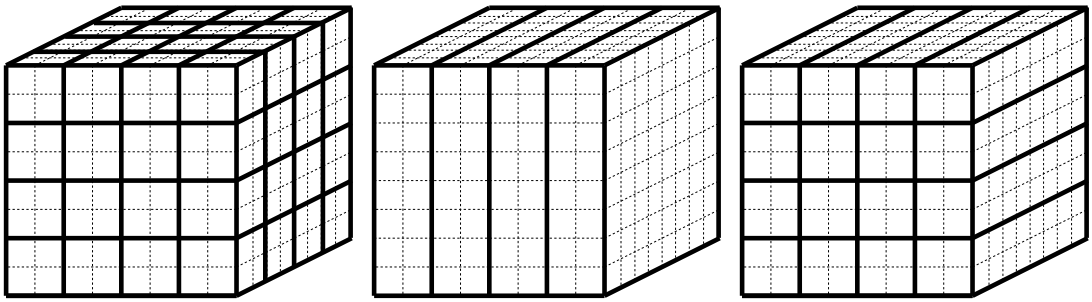
Figure 7.8: 1-D FFT with transpose. The first two stages do not involve communication. The data is then transposed among the processors. As a result, the second two stages also do not involve communication.

easily redistributed this way. The two 1-D FFTs in the plane of the slabs require no communication. The remaining 1-D FFTs require communication and could use one of the two approaches for parallel 1-D FFTs described above. A disadvantage of the slab decomposition is that for large numbers of processors, the number of processors may exceed the number of points in the 3-D FFT along any one dimension. An alternative is the pencil decomposition below.

3-D FFT with Pencil Decomposition

The pencil decomposition is shown in Fig. 7.9(c) for the case of 16 processors. Each processor holds one or more pencils of the input data. If the original input data is distributed in blocks as in Fig. 7.9(a), then communication among a row of processors (in a 3-D processor mesh) can distribute the data into the pencil decomposition. The 1-D FFTs can then be performed with no communication. To perform the 1-D FFT in another dimension, the data needs to be redistributed into pencils in another dimension. In total, four communication stages are needed for the entire 3-D FFT computation.

Victor Eijkhout



(a) Block Decomposition (b) Slab Decomposition (c) Pencil Decomposition

Figure 7.9: Three data decompositions for 3-D FFTs.

7.4 Integration for Molecular Dynamics

To numerically integrate the system of ordinary differential equations in molecular dynamics, special methods are required, different than the traditional ODE solvers that were studied in Chapter 4. These special methods, called symplectic methods, are better than other methods at producing solutions that have constant energy, for example, for systems that are called Hamiltonian (which include systems from molecular dynamics). When Hamiltonian systems are integrated with many time steps over a long time interval, preservation of structure such as total energy is often more important than the order of accuracy of the method. In this section, we motivate some ideas and give some details of the Störmer-Verlet method, which is sufficient for simple molecular dynamics simulations.

Hamiltonian systems are a class of dynamical systems which conserve energy and which can be written in a form called Hamilton's equations. Consider, for symplicity, the *simple harmonic oscillator*

$$u'' = -u$$

where u is the displacement of a single particle from an equilibrium point. This equation could model a particle with unit mass on a spring with unit spring constant. The force on a particle at position u is $-u$. This system does not look like a molecular dyanamics system but is useful for illustrating several ideas.

The above second order equation can be written as a system of first order equations

$$\begin{aligned} q' &= p \\ p' &= -q \end{aligned}$$

where $q = u$ and $p = u'$ which is common notation used in classical mechanics. The general solution is

$$\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix}.$$

The kinetic energy of the simple harmonic oscillator is $p^2/2$ and the potential energy is $q^2/2$ (the negative gradient of potential energy is the force, $-q$). Thus the total energy is proportional to $q^2 + p^2$.

Now consider the solution of the system of first order equations by three methods, explicit Euler, implicit Euler, and a method called the Störmer-Verlet method. The initial condition is $(q, p) = (1, 0)$. We use a time step of $h = 0.05$ and take 500 steps. We plot q and p on the horizontal and vertical axes, respectively (called a *phase plot*). The exact solution, as given above, is a unit circle centered at the origin.

Figure 7.10 shows the solutions. For explicit Euler, the solution spirals outward, meaning the displacement and momentum of the solution increases over time. The opposite is true for the implicit Euler method. A plot of the total energy would show the energy increasing and decreasing for the two cases, respectively. The solutions are better when smaller time steps are taken or when higher order methods are used, but these methods are not at all appropriate for integration of symplectic systems over long periods of time. Figure 7.10(c) shows the solution using a symplectic method called the Störmer-Verlet method. The solution shows that $q^2 + p^2$ is preserved much better than in the other two methods.

(a) Explicit Euler (b) Implicit Euler (c) Störmer-Verlet Method

Figure 7.10: Phase plot of the solution of the simple harmonic oscillator for three methods with initial value (1,0), time step 0.05, and 500 steps. For explicit Euler, the solution spirals outward; for implicit Euler, the solution spirals inward; the total energy is best preserved with the Störmer-Verlet method.

The Störmer-Verlet method is derived very easily. We derive it for the second order equation

$$u'' = f(t, u)$$

by simply replacing the left-hand side with a finite difference approximation

$$\frac{u_{k+1} - 2u_k + u_{k-1}}{h^2} = f(t_k, u_k)$$

which can be rearranged to obtain the method

$$u_{k+1} = 2u_k - u_{k-1} + h^2 f(t_k, u_k).$$

The formula can equivalently be derived from Taylor series. The method is similar to linear multistep methods in that some other technique is needed to supply the initial step of the method. The method is also time-reversible,

because the formula is the same if $k+1$ and $k-1$ are swapped. To explain why this method is symplectic, unfortunately, is beyond the scope of this introduction.

The method as written above has a number of disadvantages, the most severe being that the addition of the small h^2 term is subject to catastrophic cancellation. Thus this formula should not be used in this form, and a number of mathematically equivalent formulas (which can be derived from the formula above) have been developed.

One alternative formula is the *leap-frog* method:

$$\begin{aligned} u_{k+1} &= u_k + h v_{k+1/2} \\ v_{k+1/2} &= v_{k-1/2} + h f(t_k, u_k) \end{aligned}$$

where v is the first derivative (velocity) which is offset from the displacement u by a half step. This formula does not suffer from the same roundoff problems and also makes available the velocities, although they need to be re-centered with displacements in order to calculate total energy at a given step. The second of this pair of equations is basically a finite difference formula.

A third form of the Störmer-Verlet method is the *velocity Verlet* variant:

$$\begin{aligned} u_{k+1} &= u_k + h v_k + \frac{h^2}{2} f(t_k, u_k) \\ v_{k+1} &= v_k + \frac{h^2}{2} (f(t_k, u_k) + f(t_{k+1}, u_{k+1})) \end{aligned}$$

where now the velocities are computed at the same points as the displacements. Each of these algorithms can be implemented such that only two sets of quantities need to be stored (two previous positions, or a position and a velocity). These variants of the Störmer-Verlet method are popular because of their simplicity, requiring only one costly force evaluation per step. Higher-order methods have generally not been practical.

The velocity Verlet scheme is also the basis of multiple time step algorithms for molecular dynamics. In these algorithms, the slowly-varying (typically long-range) forces are evaluated less frequently and update the positions less frequently than the quickly-varying (typically short-range) forces. Finally, many state-of-the-art molecular dynamics integrate a Hamiltonian system that has been modified in order to control the simulation temperature and pressure. Much more complicated symplectic methods have been developed for these systems.

Appendix A

Theoretical background

This course requires no great mathematical sophistication. In this appendix we give a quick introduction to the few concepts that may be unfamiliar to certain readers.

A.1 Linear algebra

A.1.1 Gram-Schmidt orthogonalization

The Gram-Schmidt (GS) algorithm takes a series of vectors and inductively orthogonalizes them. This can be used to turn an arbitrary basis of a vector space into an orthogonal basis; it can also be viewed as transforming a matrix A into one with orthogonal columns. If Q has orthogonal columns, Q^tQ is diagonal, which is often a convenient property to have.

The basic principle of the GS algorithm can be demonstrated with two vectors u, v. Suppose we want a vector v' so that u, v and u, v' span the same space, but $v' \perp u$. For this we let

$$v' \leftarrow v - \frac{u^t v}{u^t u} u.$$

It is easy to see that this satisfies the requirements.

In the general case, suppose we have an orthogonal set $u_1, \ldots, u_n$ and a vector v that is not necessarily orthogonal. Then we compute

For $i = 1, \ldots, n$:
    let $c_i = u_i^t v / u_i^t u_i$
For $i = 1, \ldots, n$:
    update $v \leftarrow v - c_i u_i$

Often the vector v in the algorithm above is normalized. GS orthogonalization with this normalization, applied to a matrix, is also known as the *QR factorization*.

Exercise 1.1. Suppose that we apply the GS algorithm to the columns of a rectangular matrix A, giving a matrix Q. Prove that there is an upper triangular matrix R such that $A = QR$. If we normalize the vector v in the algorithm above, Q has the additional property that $Q^tQ = I$. Prove this too.

The GS algorithm as given above computes the desired result, but only in exact arithmetic. A computer implementation can be quite inaccurate if the angle between v and one of the u_i is small. In that case, the Modiefed Gram-Schmidt (MGS) algorithm will perform better:

For $i = 1, \ldots, n$:
    let $c_i = u_i^t v / u_i^t u_i$
    update $v \leftarrow v - c_i u_i$

To contrast it with MGS, the original GS algorithm is also known as Classical Gram-Schmidt (CGS).

As an illustration of the difference between the two methods, consider the matrix

$$A = \begin{pmatrix} 1 & 1 & 1 \\ \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon \end{pmatrix}$$

where ϵ is of the order of the machine precision, so that $1 + \epsilon^2 = 1$ in machine arithmetic. The CGS method proceeds as follows:

- The first column is of length 1 in machine arithmetic, so

$$q_1 = a_1 = \begin{pmatrix} 1 \\ \epsilon \\ 0 \\ 0 \end{pmatrix}.$$

- The second column gets orthogonalized as $v \leftarrow a_2 - 1 \cdot q_1$, giving

$$v = \begin{pmatrix} 0 \\ -\epsilon \\ \epsilon \\ 0 \end{pmatrix}, \quad \text{normalized:} \quad q_2 = \begin{pmatrix} 0 \\ -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \end{pmatrix}$$

- The third column gets orthogonalized as $v \leftarrow a_3 - c_1 q_1 - c_2 q_2$, where

$$\begin{cases} c_1 = q_1^t a_3 = 1 \\ c_2 = q_2^t a_3 = 0 \end{cases} \Rightarrow v = \begin{pmatrix} 0 \\ -\epsilon \\ 0 \\ \epsilon \end{pmatrix}; \quad \text{normalized:} \quad q_3 = \begin{pmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ 0 \\ \frac{\sqrt{2}}{2} \end{pmatrix}$$

It is easy to see that q_2 and q_3 are not orthogonal at all. By contrast, the MGS method differs in the last step:

- As before, $q_1^t a_3 = 1$, so

$$v \leftarrow a_3 - q_1 = \begin{pmatrix} 0 \\ -\epsilon \\ \epsilon \\ 0 \end{pmatrix}.$$

  Then, $q_2^t v = \frac{\sqrt{2}}{2}\epsilon$ (note that $q_2^t a_3 = 0$ before), so the second update gives

$$v \leftarrow v - \frac{\sqrt{2}}{2}\epsilon q_2 = \begin{pmatrix} 0 \\ \frac{\epsilon}{2} \\ -\frac{\epsilon}{2} \\ \epsilon \end{pmatrix}, \quad \text{normalized:} \quad \begin{pmatrix} 0 \\ \frac{\sqrt{6}}{6} \\ -\frac{\sqrt{6}}{6} \\ 2\frac{\sqrt{6}}{6} \end{pmatrix}$$

  Now all $q_i^t q_j$ are on the order of ϵ for $i \neq j$.

A.1.2 The power method

The vector sequence

$$x_i = Ax_{i-1},$$

where x_0 is some starting vector, is called the *power method* since it computes the product of subsequent matrix powers times a vector:

$$x_i = A^i x_0.$$

There are cases where the relation between the x_i vectors is simple. For instance, if x_0 is an eigenvector of A, we have for some scalar λ

$$Ax_0 = \lambda x_0 \qquad \text{and} \qquad x_i = \lambda^i x_0.$$

However, for an arbitrary vector x_0, the sequence $\{x_i\}_i$ is likely to consist of independent vectors. Up to a point.

Exercise 1.2. Let A and x be the $n \times n$ matrix and dimension n vector

$$A = \begin{pmatrix} 1 & 1 & & & \\ & 1 & 1 & & \\ & & \ddots & \ddots & \\ & & & 1 & 1 \\ & & & & 1 \end{pmatrix}, \qquad x = (0, \ldots, 0, 1)^t.$$

Show that the sequence $[x, Ax, \ldots, A^i x]$ is an independent set for $i < n$. Why is this no longer true for $i \geq n$?
Now consider the matrix B:

$$B = \left(\begin{array}{ccc|ccc} 1 & 1 & & & & \\ & \ddots & \ddots & & & \\ & & 1 & 1 & & \\ & & & 1 & & \\ \hline & & & 1 & 1 & \\ & & & & \ddots & \ddots \\ & & & & & 1 \\ & & & & & \end{array}\right), \qquad y = (0, \ldots, 0, 1)^t$$

Show that the set $[y, By, \ldots, B^i y]$ is an independent set for $i < n/2$, but not for any larger values of i.

While in general the vectors $x, Ax, A^2x, \ldots$ can be expected to be independent, in computer arithmetic this story is no longer so clear.

Suppose the matrix has eigenvalues $\lambda_0 > \lambda_1 \geq \cdots \lambda_{n-1}$ and corresponding eigenvectors u_i so that

$$Au_i = \lambda_i u_i.$$

Let the vector x be written as

$$x = c_0 u_0 + \cdots + c_{n-1} u_{n-1},$$

then

$$A^i x = c_0 \lambda_0^i u_i + \cdots + c_{n-1} \lambda_{n-1}^i u_{n-1}.$$

If we write this as

$$A^i x = \lambda_0^i \left[c_0 u_i + c_1 \left(\frac{\lambda_1}{\lambda_0} \right)^i + \cdots + c_{n-1} \left(\frac{\lambda_{n-1}}{\lambda_0} \right)^i \right],$$

we see that, *numerically*, $A^i x$ will get progressively closer to a multiple of u_0, the *dominant eigenvector*. Hence, any calculation that uses independence of the $A^i x$ vectors is likely to be inaccurate.

A.1.3 The Gershgorin theorem

Finding the eigenvalues of a matrix is usually complicated. However, there are some tools to estimate eigenvalues. In this section you will see a theorem that, in some circumstances, can give useful information on eigenvalues.

Let A be a square matrix, and x, λ an eigenpair: $Ax = \lambda x$. Looking at one component, we have

$$a_{ii} x_i + \sum_{j \neq i} a_{ij} x_j = \lambda x_i.$$

Taking norms:

$$(a_{ii} - \lambda) \leq \sum_{j \neq i} |a_{ij}| \left| \frac{x_j}{x_i} \right|$$

Taking the value of i for which $|x_i|$ is maximal, we find

$$(a_{ii} - \lambda) \leq \sum_{j \neq i} |a_{ij}|.$$

This statement can be interpreted as follows:

The eigenvalue λ is located in the circle around a_{ii} with radius $\sum_{j \neq i} |a_{ij}|$.

Since we do not know this value of i, we can only say that there is *some* value of i such that λ lies in such a circle. This is the Gershgorin theorem.

Theorem 1 *Let A be a square matrix, and let D_i be the circle with center a_{ii} and radius $\sum_{j \neq i} |a_{ij}|$, then the eigenvalues are contained in the union of circles $\cup_i D_i$.*

We can conclude that the eigenvalues are in the interior of these discs, if the constant vector is not an eigenvector.

A.2 Complexity

At various places in this book we are interested in how many operations an algorithm takes. It depends on the context what these operations are, but often we count additions (or subtractions) and multiplications. This is called the *arithmetic* or *computational complexity* of an algorithm. For instance, summing n numbers takes $n-1$ additions. Another quantity that we may want to describe is the amount of space (computer memory) that is needed. Sometimes the space to fit the input and output of an algorithm is all that is needed, but some algorithms need temporary space. The total required space is called the *space complexity* of an algorithm.

Both arithmetic and space complexity depend on some description of the input, for instance, for summing an array of numbers, the length n of the array is all that is needed. We express this dependency by saying 'summing an array of numbers has time complexity $n-1$, where n is the length of the array'.

The time (or space) the summing algorithm takes is not dependent on other factors such as the values of the numbers. By contrast, some algorithms such as computing the greatest common divisor of an array of integers *can* be dependent on the actual values.

Exercise 1.3. What is the time and space complexity of multiplying two square matrices of size $n \times n$? Assume that an addition and a multiplication take the same amount of time.

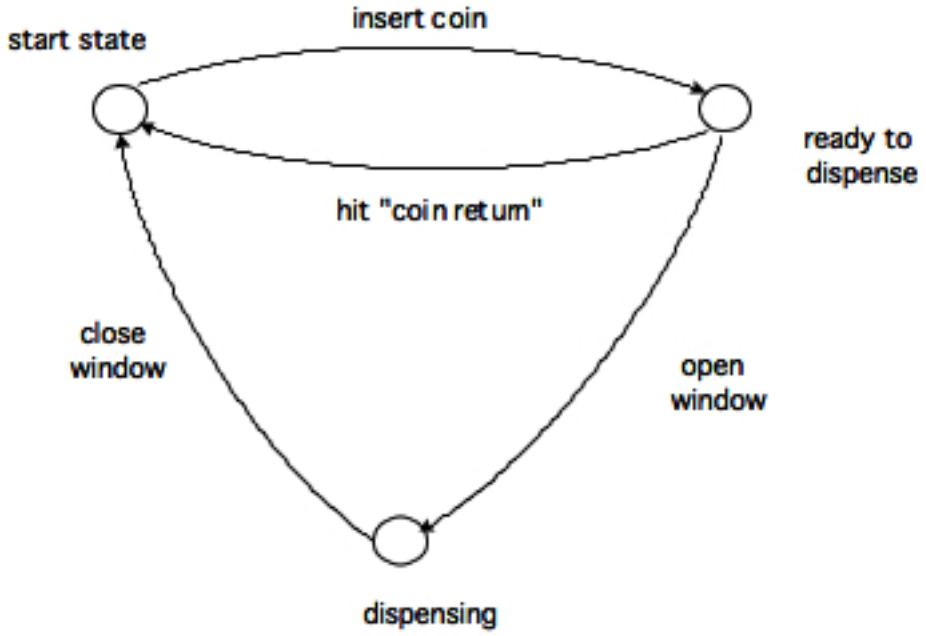
Often we aim to simplify the formulas that describe time or space complexity. For instance, if the complexity of an algorithm is $n^2 + 2n$, we see that for $n > 2$ the complexity is less than $2n^2$, and for $n > 4$ it is less than $(3/2)n^2$. On the other hand, for all values of n the complexity is at least n^2. Clearly, the quadratic term n^2 is the most important, and the linear term n becomes less and less important by ratio. We express this informally by saying that the complexity is quadratic: there are constants c, C so that for n large enough the complexity is at least cn^2 and at most Cn^2. This is expressed for short by saying that the complexity is of *order* n^2, written as $O(n^2)$.

A.3 Finite State Automatons

Finite State Automatons (FSAs) are mathematical abstractions of simple machines. For example, take a vending machine that will dispense a candy bar when a quarter has been inserted. There are four actions possible with a vending machine: insert a quarter, press 'coin return' to ask for any inserted money back, open the window to take the candy bar, and close the window again. Whether an action is possible (especially the third) depends on the *state* the machine is in. There are three states: the begin state, the state where the quarter has been inserted and the window unlocked (let us call this 'ready to dispense'), and the state where the window is open (which we will call 'dispensing').

In certain states, certain actions are not possible. For instance, in the beginning state the window can not be opened.

The mathematical description of this vending machine consists of 1. the list of states, 2. a table of how the possible actions make the machine go from one state to another. However, rather than writing down the table, a graphical representation is usually more insightful.



A.4 Partial Differential Equations

Partial Differential Equations are the source of a large fraction of HPC problems. Here is a quick derivation of two of the most important ones.

A.4.1 Partial derivatives

Derivatives of a function $u(x)$ are a measure of the rate of change. Partial derivatives to the same, but for a function $u(x,y)$ of two variables. Notated u_x and u_y, these *partial derivates* indicate the rate of change if only one variable change, and the other stays constant.

Formally, we define $u_x, u_y, u_x x$ by:

$$u_x(x,y) = \lim_{h\to 0} \frac{u(x+h,y)-u(x,y)}{h}, \quad u_y(x,y) = \lim_{h\to 0} \frac{u(x,y+h)-u(x,y)}{h}$$

A.4.2 Poisson or Laplace Equation

Let T be the temperature of a material, then its heat energy is proportional to it. A segment of length Δx has heat energy $Q = c\Delta x \cdot u$. If the heat energy in that segment is constant

$$\frac{\delta Q}{\delta t} = c\Delta x \frac{\delta u}{\delta t} = 0$$

but it is also the difference between inflow and outflow of the segment. Since flow is proportional to temperature differences, that is, to u_x, we see that also

$$0 = \left.\frac{\delta u}{\delta x}\right|_{x+\Delta x} - \left.\frac{\delta u}{\delta x}\right|_{x}$$

In the limit of $\Delta x \downarrow 0$ this gives $u_{xx} = 0$.

A.4.3 Heat Equation

Let T be the temperature of a material, then its heat energy is proportional to it. A segment of length Δx has heat energy $Q = c\Delta x \cdot u$. The rate of change in heat energy in that segment is

$$\frac{\delta Q}{\delta t} = c\Delta x \frac{\delta u}{\delta t}$$

but it is also the difference between inflow and outflow of the segment. Since flow is proportional to temperature differences, that is, to u_x, we see that also

$$\frac{\delta Q}{\delta t} = \left.\frac{\delta u}{\delta x}\right|_{x+\Delta x} - \left.\frac{\delta u}{\delta x}\right|_{x}$$

In the limit of $\Delta x \downarrow 0$ this gives $u_t = \alpha u_{xx}$.

A.4.4 Steady state

The solution of an IBVP is a function $u(x,t)$. In cases where the forcing function and the boundary conditions do not depend on time, the solution will converge in time, to a function called the *steady state* solution:

$$\lim_{t\to\infty} u(x,t) = u_{\text{steadystate}}(x).$$

This solution satisfies a BVP, which can be found by setting $u_t \equiv 0$. For instance, for the heat equation

$$u_t = u_{xx} + q(x)$$

the steady state solution satisfies $-u_{xx} = q(x)$.

A.5 Taylor series

Taylor series expansion is a powerful mathematical tool. In this course, it is used several times in proving properties of numerical methods.

The Taylor expansion theorem, in a sense, asks how well functions can be approximated by polynomials, that is, can we, for a given function f, find coefficients c_i with $i = 1, \ldots, n$ so that

$$f(x) \approx c_0 + c_1 x + c_2 x^2 + c_n x^n.$$

This question obviously needs to be refined. What do we mean by 'approximately equal'? And clearly this approximation formula can not hold for all x; the function $\sin x$ is bounded, whereas any polynomial is unbounded.

We will show that a function f with sufficiently many derivatives can be approximated as follows: if the n-th derivative $f^{(n)}$ is continuous on an interval I, then there are coefficients $c_0, \ldots, c_{n-1}$ such that

$$\forall_{x \in I} \colon |f(x) - \sum_{i<n} c_i x^i| \leq c M_n \qquad \text{where } M_n = \max_{x \in I} |f^{(n)}(x)|$$

It is easy to get inspiration for what these coefficients should be. Suppose

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots$$

(where we will not worry about matters of convergence and how long the dots go on) then filling in

$$x = 0 \text{ gives } c_0 = f(0).$$

Next, taking the first derivative

$$f'(x) = c_1 + 2c_2 x + 3c_3 x^2 + \cdots$$

and filling in

$$x = 0 \text{ gives } c_1 = f'(0).$$

From the second derivative

$$f''(x) = 2c_2 + 6c_3 x + \cdots$$

so filling in

$$x = 0 \text{ gives } c_2 = f''(0)/2.$$

Similarly, in the third derivative

$$x = 0 \text{ gives } c_3 = \tfrac{1}{3!} f^{(3)}(0).$$

Now we need to be a bit more precise. Cauchy's form of Taylor's theorem says that

$$f(x) = f(a) + \frac{1}{1!}f'(a)(x-a) + \cdots + \frac{1}{n!}f^{(n)}(a)(x-a)^n + R_n(x)$$

where the 'rest term' R_n is

$$R_n(x) = \frac{1}{(n+1)!}f^{(n+1)}(\xi)(x-a)^{n+1} \quad \text{where } \xi \in (a,x) \text{ or } \xi \in (x,a) \text{ depending.}$$

If $f^{(n+1)}$ is bounded, and $x = a + h$, then the form in which we often use Taylor's theorem is

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!}f^{(k)}(a)h^k + O(h^{n+1}).$$

We have now approximated the function f on a certain interval by a polynomial, with an error that decreases geometrically with the inverse of the degree of the polynomial.

For a proof of Taylor's theorem we use integration by parts. First we write

$$\int_a^x f'(t)dt = f(x) - f(a)$$

as

$$f(x) = f(a) + \int_a^x f'(t)dt$$

Integration by parts then gives

$$\begin{aligned} f(x) &= f(a) + [xf'(x) - af'(a)] - \textstyle\int_a^x tf''(t)dt \\ &= f(a) + [xf'(x) - xf'(a) + xf'(a) - af'(a)] - \textstyle\int_a^x tf''(t)dt \\ &= f(a) + x\textstyle\int_a^x f''(t)dt + (x-a)f'(a) - \int_a^x tf''(t)dt \\ &= f(a) + (x-a)f'(a) + \textstyle\int_a^x (x-t)f''(t)dt \end{aligned}$$

Another application of integration by parts gives

$$f(x) = f(a) + (x-a)f'(a) + \frac{1}{2}(x-a)^2 f''(a) + \frac{1}{2}\int_a^x (x-t)^2 f'''(t)dt$$

Inductively, this gives us Taylor's theorem with

$$R_{n+1}(x) = \frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)dt$$

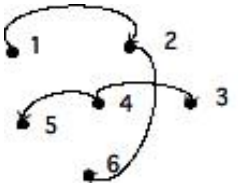
By the mean value theorem this is

$$\begin{aligned} R_{n+1}(x) &= \tfrac{1}{(n+1)!}f^{(n+1)}(\xi)\textstyle\int_a^x (x-t)^n f^{(n+1)}(t)dt \\ &= \tfrac{1}{(n+1)!}f^{(n+1)}(\xi)(x-a)^{n+1} \end{aligned}$$

A.6 Graph theory

Graph theory is the branch of mathematics that studies pairwise relations between objects. The objects, called the *nodes* or *vertices* of the graph, usually form a finite set, so we identify them with consecutive integers $1 \ldots n$ or $0 \ldots n-1$. The relation that holds between nodes is described by the *edges* of the graph: if i and j are related, we say that (i, j) is an edge of the graph.

Formally, then, a graph is a tuple $G = \langle V, E \rangle$ where $V = \{1, \ldots n\}$ for some n, and $E \subset \{(i, j) \colon 1 \leq i, j \leq n,\, i \neq j\}$.



$$\begin{cases} V = \{1, 2, 3, 4, 5, 6\} \\ E = \{(1, 2), (2, 6), (4, 3), (4, 5)\} \end{cases}$$

Figure A.1: A simple graph

A graph is called an *undirected graph* if $(i, j) \in E \Leftrightarrow (j, i) \in E$. The alternative is a *directed graph*, where we indicate an edge (i, j) with an arrow from i to j.

Two concepts that often appear in graph theory are the degree and the diameter of a graph.

Definition 1 *The* degree *denotes the maximum number of nodes that are connected to any node:*

$$d(G) \equiv \max_i |\{j \colon j \neq i \wedge (i, j) \in E\}| \,.$$

Definition 2 *The* diameter *of a graph is the length of the longest shortest path in the graph, where a path is defined as a set of vertices $v_1, \ldots, v_{k+1}$ such that $v_i \neq v_j$ for all $i \neq j$ and*

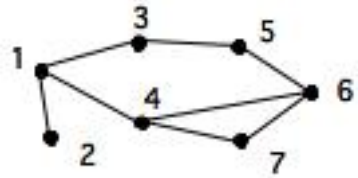
$$\forall_{1 \leq i \leq k} \colon (v_i, v_{i+1}) \in E.$$

The length of this path is k.

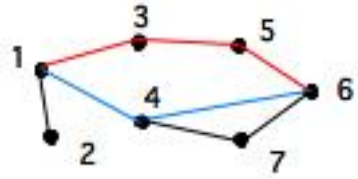
The concept of diameter is illustrated in figure A.6.

A path where all nodes are disjoint except for $v_1 = v_{k+1}$ is called a *cycle*.

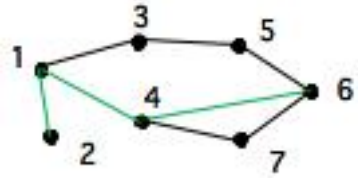
Sometimes we are only interested in the mere existence of an edge (i, j), at other times we attach a value or 'weight' w_{ij} to that edge. A graph with weighted edges is called a *weighted graph*. Such a graph can be represented as a tuple $G = \langle V, E, W \rangle$.



A graph



Two paths from 1 to 6; $\langle 1, 4, 6\rangle$ is the shorter



The longest shortest path of this graph

A.6.1 Graph colouring and independent sets

We can assign labels to the nodes of a graph, which is equivalent to partitioning the set of nodes into disjoint subsets. One type of labeling that is of interest is *graph colouring*: here the labels (or 'colours') are chosen so that, if nodes i and j have the same colour, there is no edge connecting them: $(i, j) \notin E$.

There is a trivial colouring of a graph, where each node has its own colour. More interestingly, the minimum number of colours with which you can colour a graph is called the *colour number* of the graph.

Exercise 1.4. Show that, if a graph has degree d, the colour number is at most $d + 1$.

A famous graph colouring problem is the 'four colour theorem': if a graph depicts countries on a two-dimensional map (a so-called 'planar' graph), then the colour number is at most four. In general, finding the colour number is hard (in fact, NP-hard).

The colour sets of a graph colouring are also called *independent sets*.

A.6.2 Graphs and matrices

A graph can be rendered in a number of ways. You could of course just list nodes and edges, but little insight can be derived that way. Simple graphs can be visualized by drawing vertices and edges, but for large graphs this becomes unwieldy. Another option is to construct the *adjacency matrix* of the graph. For a graph $G = \langle V, E\rangle$, the adjacency matrix M is defined by

$$n(M) = |V|, \qquad M_{ij} = \begin{cases} 1 & (i, j) \in E \\ 0 & \text{otherwise} \end{cases}$$

As an example of graph concepts that can easily be read from the adjacency matrix, consider reducibility.

Definition 3 *A graph is called* irreducible *if for every pair i, j of nodes there is a path from i to j and from j to i.*

Exercise 1.5. Let A be a matrix

$$A = \begin{pmatrix} B & C \\ \emptyset & D \end{pmatrix}$$

where B and D are square matrices. Prove the reducibility of the graph of which this is the adjacency matrix.

For graphs with edge weights, we set the elements of the adjacency matrix to the weights:

$$n(M) = |V|, \qquad M_{ij} = \begin{cases} w_{ij} & (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

You see that there is a simple correspondence between weighted graphs and matrices; given a matrix, we call the corresponding graph its *adjacency graph*.

The adjacency matrix of the graph in figure A.1 is

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

If a matrix has no zero elements, its adjacency graph has an edge between each pair of vertices. Such a graph is called a *clique*.

If the graph is undirected, the adjacency matrix is symmetric.

Here is another example of how adjacency matrices can simplify reasoning about graphs.

Exercise 1.6. Let $G = \langle V, E \rangle$ be an undirected graph, and let $G' = \langle V, E' \rangle$ be the graph with the same vertices, but with vertices defined by

$$(i,j) \in E' \Leftrightarrow \exists_k \colon (i,k) \in E \wedge (k,j) \in E.$$

If M is the adjacency matrix of G, show that M^2 is the adjacency matrix of G', where we use boolean multiplication on the elements: $1 \cdot 1 = 1, 1 + 1 + 1$.

Appendix B

Practical tutorials

This part of the book teaches you some general tools that are useful to a computational scientist. They all combine theory with a great deal of exploration. Do them while sitting at a computer!

B.1 Good coding practices

Sooner or later, and probably sooner than later, every programmer is confronted with code not behaving as it should. In this section you will learn some techniques of dealing with this problem. At first we will see a number of techniques for *preventing* errors; after that we will discuss debugging, the process of finding the, inevitable, errors in a program once they have occurred.

B.1.1 Defensive programming

In this section we will discuss a number of techniques that are aimed at preventing the likelihood of programming errors, or increasing the likelikhood of them being found at runtime. We call this *defensive programming*.

Scientific codes are often large and involved, so it is a good practice to code knowing that you are going to make mistakes and prepare for them. Another good coding practice is the use of tools: there is no point in reinventing the wheel if someone has already done it for you. Some of these tools will be described in later sections:

- Build systems, such as Make, Scons, Bjam; see section B.5.
- Source code management with SVN, Git; see section B.6.
- Regression testing and designing with testing in mind (unit testing)

First we will have a look at runtime sanity checks, where you test for things that can not or should not happen.

Example 1: if a subprogram has an array argument, it is a good idea to test whether the actual argument is a null pointer.

Example 2: if you calculate a numerical result for which certain mathematical properties hold, for instance you are writing a sine function, for which the result has to be in $[-1,1]$, you should test whether this property indeed holds for the result.

There are various ways to handle the case where the impossible happens.

B.1.1.1 The `assert` *macro*

The C standard library has a file `assert.h` which provides an `assert()` macro. Inserting `assert(foo)` has the following effect: if `foo` is zero (false), a diagnostic message is printed on standard error:

```
Assertion failed: foo, file filename, line line-number
```

which includes the literal text of the expression, the file name, and line number; and the program is subsequently aborted. Here is an example:

```
#include<assert.h>

void open_record(char *record_name)
{
  assert(record_name!=NULL);
  /* Rest of code */
```

```
}

int main(void)
{
  open_record(NULL);
}
```

The `assert` macro can be disabled by defining the `NDEBUG` macro.

B.1.1.1.1 **An assert macro for Fortran** (Thanks to Robert Mclay for this code.)

```
#if (defined( GFORTRAN ) || defined( G95 ) || defined ( PGI) )
# define MKSTR(x) "x"
#else
# define MKSTR(x) #x
#endif
#ifndef NDEBUG
# define ASSERT(x, msg) if (.not. (x) ) \
                call assert( FILE , LINE ,MKSTR(x),msg)
#else
# define ASSERT(x, msg)
#endif
subroutine assert(file, ln, testStr, msgIn)
implicit none
character(*) :: file, testStr, msgIn
integer :: ln
print *, "Assert: ",trim(testStr)," Failed at ",trim(file),":",ln
print *, "Msg:", trim(msgIn)
stop
end subroutine assert
```

B.1.1.2 Use of error codes

In some software libraries (for instance MPI or PETSc) every subprogram returns a result, either the function value or a parameter, to indicate success or failure of the routine. It is good programming practice to check these error parameters, even if you think that nothing can possibly go wrong.

It is also a good idea to write your own subprograms in such a way that they always have an error parameter. Let us consider the case of a function that performs some numerical computation.

```
float compute(float val)
{
  float result;
```

```
  result = ... /* some computation */
  return result;
}

float value,result;
result = compute(value);
```

Looks good? What if the computation can fail, for instance:

```
  result = ... sqrt(val) ... /* some computation */
```

How do we handle the case where the user passes a negative number?

```
float compute(float val)
{
  float result;
  if (val<0) { /* then what? */
  } else
    result = ... sqrt(val) ... /* some computation */
  return result;
}
```

We could print an error message and deliver some result, but the message may go unnoticed, and the calling environment does not really receive any notification that something has gone wrong.

The following approach is more flexible:

```
int compute(float val,float *result)
{
  float result;
  if (val<0) {
    return -1;
  } else {
    *result = ... sqrt(val) ... /* some computation */
  }
  return 0;
}

float value,result; int ierr;
ierr = compute(value,&result);
if (ierr!=0) { /* take appropriate action */
}
```

You can save yourself a lot of typing by writing

```
#define CHECK_FOR_ERROR(ierr) \
  if (ierr!=0) { \
```

```
    printf("Error %d detected\n",ierr); \
    return -1 ; }
....
ierr = compute(value,&result); CHECK_FOR_ERROR(ierr);
```

Using some cpp macros you can even define

```
#define CHECK_FOR_ERROR(ierr) \
  if (ierr!=0) { \
    printf("Error %d detected in line %d of file %s\n",\
           ierr,__LINE__,__FILE__); \
    return -1 ; }
```

Note that this macro not only prints an error message, but also does a further return. This means that, if you adopt this use of error codes systematically, you will get a full backtrace of the calling tree if an error occurs. (In the Python language this is precisely the wrong approach since the backtrace is built-in.)

B.1.2 Guarding against memory errors

In scientific computing it goes pretty much without saying that you will be working with large amounts of data. Some programming languages make managing data easy, other, one might say, make making errors with data easy.

The following are some examples of *memory violations*.

- Writing outside array bounds. If the address is outside the user memory, your code may abort with an error such as *segmentation violation*, and the error is reasonably easy to find. If the address is just outside an array, it will corrupt data; such an error may go undetected for a long time, as it can have no effect, or only introduce subtly wrong values in your computation.
- Reading outside array bounds can be harder to find than errors in writing, as it will often not abort your code, but only introduce wrong values.
- The use of uninitialized memory is similar to reading outside array bounds, and can go undetected for a long time. One variant of this is through attaching memory to an unallocated pointer.
  This particular kind of error can manifest itself in interesting behaviour. Let's say you notice that your program misbehaves, you recompile it with debug mode to find the error, and now the error no longer occurs. This is probably due to the effect that, with low optimization levels, all allocated arrays are filled with zeros. Therefore, your code was originally reading a random value, but is now getting a zero.

This section contains some techniques to prevent errors in dealing with memory that you have reserved for your data.

B.1.2.1 Array bound checking and other memory techniques

In parallel codes, memory errors will often show up by a crash in an MPI routine. This is hardly ever an MPI problem or a problem with your cluster.

Compilers for Fortran often have support for array bound checking. Since this makes your code much slower, you would only enable it during the development phase of your code.

B.1.2.2 Memory leaks

We say that a program has a *memory leak*, if it allocates memory, and subsequently loses track of that memory. The operating system then thinks the memory is in use, while it is not, and as a result the computer memory can get filled up with allocated memory that serves no useful purpose.

Example:

```
for (i=.... ) {
  real *block = malloc( /* large number of bytes */ )
  /* do something with that block of memory */
  /* and forget to call "free" on that block */
}
```

There are various tools for detecting memory errors: Valgrind, DMALLOC, Electric Fence.

Valgrind is fairly easy to use. After you compile your program, you call

```
valgrind yourprogram # with options, if needed
```

and valgrind will give you its diagnosis. For example, the following program reads out of bounds: and valgrind reports this:

```
 %% valgrind bounds
==2707== Memcheck, a memory error detector.
==2707== Copyright (C) 2002-2005, and GNU GPL'd, by Julian Seward et al.
==2707== Using LibVEX rev 1575, a library for dynamic binary translation.
==2707== Copyright (C) 2004-2005, and GNU GPL'd, by OpenWorks LLP.
==2707== Using valgrind-3.1.1, a dynamic binary instrumentation framework.
==2707== Copyright (C) 2000-2005, and GNU GPL'd, by Julian Seward et al.
==2707== For more details, rerun with: -v
==2707==
==2707== Invalid read of size 4
==2707==    at 0x400A24: main (in /home/eijkhout/valgrind/bounds)
==2707==  Address 0x4A23044 is 0 bytes after a block of size 20 alloc'd
==2707==    at 0x4904A06: malloc (vg_replace_malloc.c:149)
==2707==    by 0x400A20: main (in /home/eijkhout/valgrind/bounds)
surprise: 0.000000e+00
==2707==
==2707== ERROR SUMMARY: 1 errors from 1 contexts (suppressed: 5 from 2)
==2707== malloc/free: in use at exit: 0 bytes in 0 blocks.
==2707== malloc/free: 1 allocs, 1 frees, 20 bytes allocated.
==2707== For counts of detected errors, rerun with: -v
==2707== All heap blocks were freed -- no leaks are possible.
```

You see that the output does not refer to any program lines or variables, since valgrind is working with the binary here. If you compile your code with `-g`, you get at least the line numbers:

```
==2733== Invalid read of size 4
==2733==    at 0x40058A: main (bounds.c:6)
==2733==  Address 0x4A23044 is 0 bytes after a block of size 20 alloc'd
==2733==    at 0x4904A06: malloc (vg_replace_malloc.c:149)
==2733==    by 0x400579: main (bounds.c:5)
```

It should be noted that the PETSc library can not catch all errors that valgrind does, but on the ones it does it gives more informative error messages.

B.1.2.3 Roll-your-own malloc

Many programming errors arise from improper use of dynamically allocated memory: the program writes beyond the bounds, or writes to memory that has not been allocated yet, or has already been freed. While some compilers can do bound checking at runtime, this slows down your program. A better strategy is to write your own memory management.

If you write in C, you are surely familiar with the `malloc` and `free` calls:

```
int *ip;
ip = (int*) malloc(500*sizeof(int));
if (ip==0) {/* could not allocate memory */}
..... do stuff with ip .....
free(ip);
```

You can save yourself some typing by

```
#define MYMALLOC(a,b,c) \
  a = (c*)malloc(b*sizeof(c)); \
  if (a==0) {/* error message and appropriate action */}

int *ip;
MYMALLOC(ip,500,int);
```

Runtime checks on memory usage (either by compiler-generated bounds checking, or through tools like valgrind or Rational Purify) are expensive, but you can catch many problems by adding some functionality to your malloc. What we will do here is to detect memory corruption after the fact.

We allocate a few integers to the left and right of the allocated object (line `1` in the code below), and put a recognizable value in them (line `2` and `3`), as well as the size of the object (line `2`). We then return the pointer to the actually requested memory area (line `4`).

```
#define MEMCOOKIE 137
#define MYMALLOC(a,b,c) { \
  char *aa; int *ii; \
  aa = malloc(b*sizeof(c)+3*sizeof(int)); /* 1 */ \
  ii = (int*)aa; ii[0] = b*sizeof(c); \
          ii[1] = MEMCOOKIE;              /* 2 */ \
```

```
  aa = (char*)(ii+2); a = (c*)aa ;         /* 4 */ \
  aa = aa+b*sizesof(c); ii = (int*)aa; \
          ii[0] = MEMCOOKIE;               /* 3 */ \
  }
```

Now you can write your own `free`, which tests whether the bounds of the object have not been written over.

```
#define MYFREE(a) { \
  char *aa; int *ii,; ii = (int*)a; \
  if (*(--ii)!=MEMCOOKIE) printf("object corrupted\n"); \
  n = *(--ii); aa = a+n; ii = (int*)aa; \
  if (*ii!=MEMCOOKIE)  printf("object corrupted\n"); \
  }
```

You can extend this idea: in every allocated object, also store two pointers, so that the allocated memory areas become a doubly linked list. You can then write a macro `CHECKMEMORY` which tests all your allocated objects for corruption.

Such solutions to the memory corruption problem are fairly easy to write, and they carry little overhead. There is a memory overhead of at most 5 integers per object, and there is absolutely no performance penalty.

(Instead of writing a wrapper for `malloc`, on some systems you can influence the behaviour of the system routine. On linux, `malloc` calls hooks that can be replaced with your own routines; see `http://www.gnu.org/s/libc/manual/html_node/Hooks-for-Malloc.html`.)

B.1.2.4 Specific techniques: Fortran

Use `Implicit none`.

Put all subprograms in modules so that the compiler can check for missing arguments and type mismatches. It also allows for automatic dependency building with `fdepend`.

Use the C preprocessor for conditional compilation and such.

B.1.3 Debugging

When a program misbehaves, *debugging* is the process of finding out *why*. There are various strategies of finding errors in a program. The crudest one is debugging by *print statements*. If you have a notion of where in your code the error arises, you can edit your code to insert print statements, recompile, rerun, and see if the output gives you any suggestions. There are several problems with this:

- The edit/compile/run cycle is time consuming, especially since
- often the error will be caused by an earlier section of code, requiring you to edit and rerun. Furthermore,
- the amount of data produced by your program can be too large to display and inspect effectively.
- If your program is parallel, you probably need to print out data from all proccessors, making the inspection process very tedious.

For these reasons, the best way to debug is by the use of an interactive *debugger*, a program that allows you to monitor and control the behaviour of a running program. In this section you will familiarize yourself with *gdb*, which is the open source debugger of the *GNU* project. Other debuggers are proprietary, and typically come with a compiler suite. Another distinction is that `gdb` is a commandline debugger; there are graphical debuggers such as *ddd* (a frontend to `gdb`) or *DDT* and *TotalView* (debuggers for parallel codes). We limit ourselves to `gdb`, since it incorporates the basic concepts common to all debuggers.

B.1.3.1 Invoking `gdb`

There are three ways of using `gdb`: using it to start a program, attaching it to an already running program, or using it to inspect a *core dump*. We will only consider the first possibility.

Here is an exaple of how to start `gdb` with program that has no arguments:

tutorials/gdb/hello.c

```
#include <stdlib.h>
#include <stdio.h>
int main() {
  printf("hello world\n");
  return 0;
}
```

```
%% cc -g -o hello hello.c
# regular invocation:
%% ./hello
hello world
# invocation from gdg:
%% gdb hello
GNU gdb 6.3.50-20050815 # ..... version info
Copyright 2004 Free Software Foundation, Inc. .... copyright info ....
(gdb) run
Starting program: /home/eijkhout/tutorials/gdb/hello
Reading symbols for shared libraries +. done
hello world

Program exited normally.
(gdb) quit
%%
```

Important note: the program was compiled with the *debug flag* `-g`. This causes the 'symbol table' and other debug information to be included in the binary. This will make your binary larger than strictly necessary, but it will also make it slower, for instance because the compiler will not perform certain optimizations[1].

1. Compiler optimizations are not supposed to change the semantics of a program, but sometimes do. This can lead to the nightmare scenario where a program crashes or gives incorrect results, but magically works correctly with compiled with debug and run in a debugger.

For a program with commandline input we give the arguments to the `run` command:

tutorials/gdb/say.c

```
#include <stdlib.h>
#include <stdio.h>
int main(int argc,char **argv) {
  int i;
  for (i=0; i<atoi(argv[1]); i++)
    printf("hello world\n");
  return 0;
}
```

```
%% ./say 2
hello world
hello world
%% gdb say
.... the usual messages ...
(gdb) run 2
Starting program: /home/eijkhout/tutorials/gdb/say 2
Reading symbols for shared libraries +. done
hello world
hello world

Program exited normally.
```

Let us now consider a program with several errors:

tutorials/gdb/square.c

```
#include <stdlib.h>
#include <stdio.h>
int main(int argc,char **argv) {
  int nmax,i;
  float *squares,sum;

  fscanf(stdin,"%d",nmax);
  for (i=1; i<=nmax; i++) {
    squares[i] = 1./(i*i); sum += squares[i];
  }
  printf("Sum: %e\n",sum);

  return 0;
}
```

```
%% cc -g -o square square.c
 %% ./square
50
```

```
Segmentation fault
```

The *segmentation fault* indicates that we are accessing memory that we are not allowed to, making the program abort. A debugger will quickly tell us where this happens:

```
[albook:~/Current/istc/scicompbook/tutorials/gdb] %% gdb square
(gdb) run
50000

Program received signal EXC_BAD_ACCESS, Could not access memory.
Reason: KERN_INVALID_ADDRESS at address: 0x000000000000eb4a
0x00007fff824295ca in __svfscanf_l ()
```

Apparently the error occurred in a function `__svfscanf_l`, which is not one of ours, but a system function. Using the `backtrace` (or `bt`, also `where` or `w`) command we quickly find out how this came to be called:

```
(gdb) backtrace
#0  0x00007fff824295ca in __svfscanf_l ()
#1  0x00007fff8244011b in fscanf ()
#2  0x0000000100000e89 in main (argc=1, argv=0x7fff5fbfc7c0) at square.c:7
```

We take a close look at line 7, and change `nmax` to `&nmax`.

There is still an error in our program:

```
(gdb) run
50000

Program received signal EXC_BAD_ACCESS, Could not access memory.
Reason: KERN_PROTECTION_FAILURE at address: 0x000000010000f000
0x0000000100000ebe in main (argc=2, argv=0x7fff5fbfc7a8) at square1.c:9
9           squares[i] = 1./(i*i); sum += squares[i];
```

We investigate further:

```
(gdb) print i
$1 = 11237
(gdb) print squares[i]
Cannot access memory at address 0x10000f000
```

and we quickly see that we forgot to allocate `squares`!

By the way, we were lucky here: this sort of memory errors is not always detected. Starting our programm with a smaller input does not lead to an error:

```
(gdb) run
50
Sum: 1.625133e+00

Program exited normally.
```

Victor Eijkhout

For further study: `http://heather.cs.ucdavis.edu/~matloff/UnixAndC/CLanguage/Debug.html`.

B.1.4 Testing

There are various philosophies for testing the correctness of a code.

- Correctness proving: the programmer draws up predicates that describe the intended behaviour of code fragments and proves by mathematical techniques that these predicates hold [52, 24].
- Unit testing: each routine is tested separately for correctness. This approach is often hard to do for numerical codes, since with floating point numbers there is essentially an infinity of possible inputs, and it is not easy to decide what would constitute a sufficient set of inputs.
- Integration testing: test subsystems
- System testing: test the whole code. This is often appropriate for numerical codes, since we often have model problems with known solutions, or there are properties such as bounds that need to hold on the global solution.
- Test-driven design

With parallel codes we run into a new category of difficulties with testing. Many algorithms, when executed in parallel, will execute operations in a slightly different order, leading to different roundoff behaviour. For instance, the parallel computation of a vector sum will use partial sums. Some algorithms have an inherent damping of numerical errors, for instance stationary iterative methods (section 5.5.1), but others have no such built-in error correction (nonstationary methods; section 5.5.7). As a result, the same iterative process can take different numbers of iterations depending on how many processors are used.

B.1.4.1 Test-driven design and development

- Always have a testable code.
- Make only the smallest change that allows for testing.
- With every change, test before and after.
- Assure correctness before adding new features.

B.2 LaTeX for scientific documentation

B.2.1 The idea behind LaTeX, some history

TeX is a typesetting system that dates back to the late 1970s. In those days, graphics terminals were you could design a document layout and immediately view it, the way you can with Microsoft Word, were rare. Instead, TeX uses a two-step workflow, where you first type in your document with formatting instructions in an ascii document, using your favourite text editor. Next, you would invoke the `latex` program, as a sort of compiler, to translate this document to a form that can be printed or viewed.

```
%% edit mydocument.tex
%% latex mydocument
%% # print or view the resulting output
```

The process is comparable to making web pages by typing HTML commands.

This way of working may seem clumsy, but it has some advantages. For instance, the TeX input files are plain ascii, so they can easily be generated automatically, for instance from a database. Also, you can edit them with whatever your favourite editor happens to be.

Another point in favour of TeX is the fact that the layout is specified by commands that are written in a sort of programming language. This has some important consequences:

- Separation of concerns: when you are writing your document, you do not have to think about layout. You give the 'chapter' command, and the implementation of that command will be decided independently, for instance by you choosing a document style.
- Changing the layout of a finished document is easily done by choosing a different realization of the layout commands in the input file: the same 'chapter' command is used, but by choosing a different style the resulting layout is different. This sort of change can be as simple as a one-line change to the document style declaration.
- If you have unusual typesetting needs, it is possible to write new TeX commands for this. For many needs such extensions have in fact already been written; see section B.2.4.

The commands in TeX are fairly low level. For this reason, a number of people have written systems on top of TeX that offer powerful features, such as automatic cross-referencing, or generation of a table of contents. The most popular of these systems is LaTeX. Since TeX is an interpreted system, all of its mechanisms are still available to the user, even though LaTeX is loaded on top of it.

B.2.1.1 Installing LaTeX

The easiest way to install LaTeX on your system is by downloading the TeXlive distribution from `http://tug.org/texlive`. Apple users can also use `fink` or `macports`. Various front-ends to TeX exist, such as TeXshop on the Mac.

B.2.1.2 Running LaTeX

Purpose. In this section you will run the LaTeX compiler

Originally, the `latex` compiler would output a device independent file format, named `dvi`, which could then be translated to PostScript or PDF, or directly printed. These days, many people use the `pdflatex` program which directly translates `.tex` files to `.pdf` files. This has the big advantage that the generated PDF files have automatic cross linking and a side panel with table of contents. An illustration is found below.

Let us do a simple example.

```
\documentclass{article}
\begin{document}
Hello world!
\end{document}
```

Figure B.1: A minimal LaTeX document

Exercise. Create a text file `minimal.tex` with the content as in figure B.1. Try the command `pdflatex minimal` or `latex minimal`. Did you get a file `minimal.pdf` in the first case or `minimal.dvi` in the second case? Use a pdf viewer, such as Adobe Reader, or `dvips` respectively to view the output.

Caveats. If you make a typo, TeX can be somewhat unfriendly. If you get an error message and TeX is asking for input, typing `x` usually gets you out, or `Ctrl-C`. Some systems allow you to type `e` to go directly into the editor to correct the typo.

B.2.2 A gentle introduction to LaTeX

Here you will get a very brief run-through of LaTeX features. There are various more in-depth tutorials available, such as the one by Oetiker [65].

B.2.2.1 Document structure

Each LaTeX document needs the following lines:

```
\documentclass{ .... } % the dots will be replaced

\begin{document}

\end{document}
```

The 'documentclass' line needs a class name in between the braces; typical values are 'article' or 'book'. Some organizations have their own styles, for instance 'ieeeproc' is for proceedings of the IEEE.

All document text goes between the `\begin{document}` and `\end{document}` lines. (Matched 'begin' and 'end' lines are said to denote an 'environment', in this case the document environment.)

The part before `\begin{document}` is called the 'preamble'. It contains customizations for this particular document. For instance, a command to make the whole document double spaced would go in the preamble. If you are using `pdflatex` to format your document, you want a line

```
\usepackage{hyperref}
```

here.

Have you noticed the following?

- The backslash character is special: it starts a LaTeX command.
- The braces are also special: they have various functions, such as indicating the argument of a command.
- The percent character indicates that everything to the end of the line is a comment.

B.2.2.2 Some simple text

Purpose. In this section you will learn some basics of text formatting.

Exercise. Create a file `first.tex` with the content of figure B.1 in it. Type some text before the `\begin{document}` line and run `pdflatex` on your file.

Expected outcome. You should get an error message: you are not allowed to have text before the `\begin{document}` line. Only commands are allowed in the preamble part.

Exercise. Edit your document: put some text in between the `\begin{document}` and `\end{document}` lines. Let your text have both some long lines that go on for a while, and some short ones. Put superfluous spaces between words, and at the beginning or end of lines. Run `pdflatex` on your document and view the output.

Expected outcome. You notice that the white space in your input has been collapsed in the output. TeX has its own notions about what space should look like, and you do not have to concern yourself with this matter.

Exercise. Edit your document again, cutting and pasting the paragraph, but leaving a blank line between the two copies. Paste it a third time, leaving several blank lines. Format, and view the output.

Expected outcome. TeX interprets one or more blank lines as the separation between paragraphs.

Exercise. Add `\usepackage{pslatex}` to the preamble and rerun `pdflatex` on your document. What changed in the output?

Expected outcome. This should have the effect of changing the typeface from the default to Times Roman.

Caveats. Typefaces are notoriously unstandardized. Attempts to use different typefaces may or may not work. Little can be said about this in general.

Add the following line before the first paragraph:

```
\section{This is a section}
```

and a similar line before the second. Format. You see that LaTeX automatically numbers the sections, and that it handles indentation different for the first paragraph after a heading.

Exercise. Replace `article` by `artikel3` in the documentclass declaration line and reformat your document. What changed?

Expected outcome. There are many documentclasses that implement the same commands as `article` (or another standard style), but that have their own layout. Your document should format without any problem, but get a better looking layout.

Caveats. The `artikel3` class is part of most distributions these days, but you can get an error message about an unknown documentclass if it is missing or if your environment is not set up correctly. This depends on your installation. If the file seems missing, download the files from `http://tug.org/texmf-dist/tex/latex/ntgclass/` and put them in your current directory; see also section B.2.2.8.

B.2.2.3 Math

Purpose. In this section you will learn the basics of math typesetting

One of the goals of the original TeX system was to facilitate the setting of mathematics. There are two ways to have math in your document:

- Inline math is part of a paragraph, and is delimited by dollar signs.
- Display math is, as the name implies, displayed by itself.

Exercise. Put `$x+y$` somewhere in a paragraph and format your document. Put `\[x+y\]` somewhere in a paragraph and format.

Expected outcome. Formulas between single dollars are included in the paragraph where you declare them. Formulas between `\[...\]` are typeset in a display.

For display equations with a number, use an `equation` environment. Try this.

Here are some common things to do in math. Make sure to try them out.

- Subscripts and superscripts: `$x_i^2$`. If the sub or superscript is more than a single symbol, it needs to be grouped: `$x_{i+1}^{2n}$`. (If you need a brace in a formula, use `$\{ \}$`.
- Greek letters and other symbols: `$\alpha\otimes\beta_i$`.
- Combinations of all these `$\int_{t=0}^\infty tdt$`.

Exercise. Take the last example and typeset it as display math. Do you see a difference with inline math?

Expected outcome. TeX tries not to include the distance between text lines, even if there is math in a paragraph. For this reason it typesets the bounds on an integral sign differently from display math.

B.2.2.4 Referencing

Purpose. In this section you will see TeX's cross referencing mechanism in action.

So far you have not seen LaTeX do much that would save you any work. The cross referencing mechanism of LaTeX will definitely save you work: any counter that LaTeX inserts (such as section numbers) can be referenced by a label. As a result, the reference will always be correct.

Start with an example document that has at least two section headings. After your first section heading, put the command `\label{sec:first}`, and put `\label{sec:other}` after the second section heading. These label commands can go on the same line as the `section` command, or on the next. Now put

```
As we will see in section~\ref{sec:other}.
```

in the paragraph before the second section. (The tilde character denotes a non-breaking space.)

Exercise. Make these edits and format the document. Do you see the warning about an undefined reference? Take a look at the output file. Format the document again, and check the output again. Do you have any new files in your directory?

Expected outcome. On a first pass through a document, the TeX compiler will gather all labels with their values in a `.aux` file. The document will display a double questioni mark for any references that are unknown. In the second pass the correct values will be filled in.

Caveats. If after the second pass there are still undefined references, you probably made a typo. If you use the `bibtex` utility for literature references, you will regularly need three passes to get all references resolved correctly.

Above you saw that the `equation` environment gives displayed math with an equation number. You can add a label to this environment to refer to the equation number.

Exercise. Write a formula in an `equation` environment, and add a label. Refer to this label anywhere in the text. Format (twice) and check the output.

Expected outcome. The `\label` and `\ref` command are used in the same way for formulas as for section numbers. Note that you must use `\begin/end{equation}` rather than `\[...\]` for the formula.

B.2.2.5 Lists

Purpose. In this section you will see the basics of lists.

Bulleted and numbered lists are provided through an environment.

```
\begin{itemize}
\item This is an item;
\item this is one too.
```

```
\end{itemize}
\begin{enumerate}
\item This item is numbered;
\item this one is two.
\end{enumerate}
```

Exercise. Add some lists to your document, including nested lists. Inspect the output.

Expected outcome. Nested lists will be indented further and the labeling and numbering style changes with the list depth.

Exercise. Add a label to an item in an `enumerate` list and refer to it.

Expected outcome. Again, the `\label` and `\ref` commands work as before.

B.2.2.6 Graphics

Since you can not immediately see the output of what you are typing, sometimes the output may come as a surprise. That is especially so with graphics. LaTeX has no standard way of dealing with graphics, but the following is a common set of commands:

```
\usepackage{graphicx} % this line in the preamble

\includegraphics{myfigure} % in the body of the document
```

The figure can be in any of a number of formats, except that PostScript figures (with extension `.ps` or `.eps`) can not be used if you use pdflatex.

Since your figure is often not the right size, the include line will usually have something like:

```
\includegraphics[scale=.5]{myfigure}
```

A bigger problem is that figures can be too big to fit on the page if they are placed where you declare them. For this reason, they are usually treated as 'floating material'. Here is a typical declaration of a figure:

```
\begin{figure}[ht]
  \includegraphics{myfigure}
  \caption{This is a figure}
  \label{fig:first}
\end{figure}
```

It contains the following elements:

- The `figure` environment is for 'floating' figures; they can be placed right at the location where they are declared, at the top or bottom of the next page, at the end of the chapter, et cetera.
- The `[ht]` argument of the `\begin{figure}` line states that your figure should be attempted to be placed `h`ere; it that does not work, it should go `t`op of the next page. The remaining possible specifications are `b` for placement at the bottom of a page, or `p` for placement on a page by itself. For example

```
\begin{figure}[hbp]
```
declares that the figure has to be placed here if possible, at the bottom of the next page if that's not possible, and on a page of its own if it is too big to fit on a page with text.

- A caption to be put under the figure, including a figure number;
- A label so that you can refer to the figure by `as can be seen in figure~\ref{fig:first}`.
- And of course the figure material. There are various ways to fine-tune the figure placement. For instance
```
\begin{center}
  \includegraphics{myfigure}
\end{center}
```
gives a centered figure.

B.2.2.7 Bibliography references

The mechanism for citing papers and books in your document is a bit like that for cross referencing. There are labels involved, and there is a `\cite{thatbook}` command that inserts a reference, usually numeric. However, since you are likely to refer to a paper or book in more than one document your write, LaTeX allows you to have a database of literature references in a file by itself, rather than somewhere in your document.

Make a file `mybibliography.bib` with the following content:

```
@article{JoeDoe1985,
author = {Joe Doe},
title = {A framework for bibliography references},
journal = {American Library Assoc. Mag.},
year = {1985}
}
```

In your document `mydocument.tex`, put

```
For details, refer to Doe~\cite{JoeDoe1985} % somewhere in the text

\bibliography{mybibliography} % at the end of the document
\bibliographystyle{plain}
```

Format your document, then type on the commandline

```
bibtex mydocument
```

and format your document two more times. There should now be a bibliography in it, and a correct citation. You will also see that files `mydocument.bbl` and `mydocument.blg` have been created.

B.2.2.8 Styles and other customizations

On Unix systems, TeX investigates the `TEXINPUTS` environment variable when it tries to find an include file. Consequently, you can create a directory for your styles and other downloaded include files, and set this variable to the location of that directory.

Victor Eijkhout

B.2.3 A worked out example

The following example `demo.tex` contains many of the elements discussed above.

```
\documentclass{artikel3}

\usepackage{pslatex}
\usepackage{amsmath,amssymb}
\usepackage{graphicx}

\newtheorem{theorem}{Theorem}

\newcounter{excounter}
\newenvironment{exercise}
  {\refstepcounter{excounter}
   \begin{quotation}\textbf{Exercise \arabic{excounter}.} }
  {\end{quotation}}

\begin{document}
\title{SSC 335: demo}
\author{Victor Eijkhout}
\date{today}
\maketitle

\section{This is a section}
\label{sec:intro}

This is a test document, used in~\cite{latexdemo}. It contains a
discussion in section~\ref{sec:discussion}.

\begin{exercise}\label{easy-ex}
  Left to the reader.
\end{exercise}
\begin{exercise}
  Also left to the reader, just like in exercise~\ref{easy-ex}
\end{exercise}

\begin{theorem}
  This is cool.
\end{theorem}
This is a formula: $a\Leftarrow b$.
\begin{equation}
  \label{eq:one}
    x_i\leftarrow y_{ij}\cdot x^{(k)}_j
\end{equation}
Text: $\int_0^1 \sqrt x\,dx$
\[
```

```
  \int_0^1 \sqrt x\,dx
\]
\section{This is another section}
\label{sec:discussion}

\begin{table}[ht]
  \centering
  \begin{tabular}{|rl|}
    \hline one&value \\ \hline another&values \\ \hline
  \end{tabular}
  \caption{This is the only table in my demo}
  \label{tab:thetable}
\end{table}
\begin{figure}[ht]
  \centering
  \includegraphics{graphics/caches}
  \caption{this is the only figure}
  \label{fig:thefigure}
\end{figure}
As I showed in the introductory section~\ref{sec:intro}, in the
paper~\cite{AdJo:colorblind}, it was shown that
equation~\eqref{eq:one}
\begin{itemize}
\item There is an item.
\item There is another item
  \begin{itemize}
  \item sub one
  \item sub two
  \end{itemize}
\end{itemize}
\begin{enumerate}
\item item one
\item item two
  \begin{enumerate}
  \item sub one
  \item sub two
  \end{enumerate}
\end{enumerate}

\tableofcontents
\listoffigures

\bibliography{math}
\bibliographystyle{plain}

\end{document}
```

Victor Eijkhout

You also need the file `math.bib`:

```
@article{AdJo:colorblind,
author = {Loyce M. Adams and Harry F. Jordan},
title = {Is {SOR} color-blind?},
journal = {SIAM J. Sci. Stat. Comput.},
year = {1986},
volume = {7},
pages = {490--506},
abstract = {For what stencils do ordinary and multi-colour SOR have
the same eigenvalues.},
keywords = {SOR, colouring}
}

@misc{latexdemo,
author = {Victor Eijkhout},
title = {Short {\LaTeX}\ demo},
note = {SSC 335, oct 1, 2008}
}
```

The following sequence of commands

```
pdflatex demo
bibtex demo
pdflatex demo
pdflatex demo
```

gives

SSC 335: demo

Victor Eijkhout

today

1 This is a section

This is a test document, used in [2]. It contains a discussion in section 2.

Exercise 1. Left to the reader.

Exercise 2. Also left to the reader, just like in exercise 1

Theorem 1 *This is cool.*

This is a formula: $a \Leftarrow b$.

$$x_i \leftarrow y_{ij} \cdot x_j^{(k)} \tag{1}$$

Text: $\int_0^1 \sqrt{x}dx$

$$\int_0^1 \sqrt{x}dx$$

2 This is another section

one	value
another	values

Table 1: This is the only table in my demo

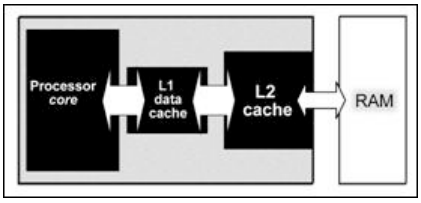


Figure 1: this is the only figure

As I showed in the introductory section 1, in the paper [1], it was shown that equation (1)

- There is an item.

- There is another item
  - sub one
  - sub two

1. item one
2. item two
   (a) sub one
   (b) sub two

Contents

List of Figures

References

[1] Loyce M. Adams and Harry F. Jordan. Is SOR color-blind? *SIAM J. Sci. Stat. Comput.*, 7:490–506, 1986.
[2] Victor Eijkhout. Short LaTeX demo. SSC 335, oct 1, 2008.

B.2.4 Where to take it from here

This tutorial touched only briefly on some essentials of TeX and LaTeX. You can find longer intros online [65], or read a book [58, 56, 64]. Macro packages and other software can be found on the Comprehensive TeX Archive `http://www.ctan.org`. For questions you can go to the newsgroup `comp.text.tex`, but the most common ones can often already be found on web sites [77].

Victor Eijkhout

B.3 Unix intro

Unix is an *Operating System (OS)*, that is, a layer of software between the user or a user program and the hardware. It takes care of files, screen output, and it makes sure that many processes can exist side by side on one system. However, it is not immediately visible to the user. Most of the time that you use Unix, you are typing commands which are executed by an interpreter called the *shell*. The shell makes the actual OS calls. There are a few possible Unix shells available, but in this tutorial we will assume that you are using the *sh* or *bash* shell, although many commands are common to the various shells in existence.

This short tutorial will get you going; if you want to learn more about Unix and shell scripting, see for instance `http://www.tldp.org/guides.html`.

B.3.1 Files and such

Purpose. In this section you will learn about the Unix file system, which consists of *directories* that store *files*. You will learn about *executable* files and commands for displaying data files.

B.3.1.1 Looking at files

Purpose. In this section you will learn commands for displaying file contents.

The `ls` command gives you a listing of files that are in your present location.

Exercise. Type `ls`. Does anything show up?

Expected outcome. If there are files in your directory, they will be `ls`ted; if there are none, no output will be given. This is standard Unix behaviour: no output does not mean that something went wrong, it only means that there is nothing to report.

The `cat` command is often used to display files, but it can also be used to create some simple content.

Exercise. Type `cat > newfilename` (where you can pick any filename) and type some text. Conclude with `Control-d` on a line by itself. Now use `cat` to view the contents of that file: `cat newfilename`.

Expected outcome. In the first use of `cat`, text was con`cat`enated from the terminal to a file; in the second the file was cat'ed to the terminal output. You should see on your screen precisely what you typed into the file.

Caveats. Be sure to type `Control-d` as the first thing on the last line of input. If you really get stuck, `Control-c` will usually get you out. Try this: start creating a file with `cat > filename` and hit `Control-c` in the middle of a line. What are the contents of your file?

Above you used `ls` to get a directory listing. You can also use the `ls` command on specific files:

Exercise. Do `ls newfilename` with the file that you created above; also do `ls nosuchfile` with a file name that does not exist.

Expected outcome. For an existing file you get the file name on your screen; for a non-existing file you get an error message.

The `ls` command can give you all sorts of information.

Exercise. Read the man page of the `ls` command: `man ls`. Find out the size and creation date of some files, for instance the file you just created.

Expected outcome. Did you find the `ls -s` and `ls -l` options? The first one lists the size of each file, usually in kilobytes, the other gives all sorts of information about a file, including things you will learn about later.

Caveats. The `man` command puts you in a mode where you can view long text documents. This viewer is common on Unix systems (it is available as the `more` or `less` system command), so memorize the following ways of navigating: Use the space bar to go forward and the `u` key to go back up. Use `g` to go to the beginning fo the text, and `G` for the end. Use `q` to exit the viewer. If you really get stuck, `Control-c` will get you out.

(If you already know what command you're looking for, you can use man to get online information about it. If you forget the name of a command, `man -k keyword` can help you find it.)

Three more useful commands for files are: `cp` for copying, `mv` (short for 'move') for renaming, and `rm` ('remove') for deleting. Experiment with them.

There are more commands for displaying a file, parts of a file, or information about a file.

Exercise. Do `ls /usr/share/words` or `ls /usr/share/dict/words` to confirm that a file with words exists on your system. Now experiment with the commands `head`, `tail`, `more`, and `wc` using that file.

Expected outcome. `head` displays the first couple of lines of a file, `tail` the last, and `more` uses the same viewer that is used for man pages. The `wc` ('word count') command reports the number of words, characters, and lines in a file.

B.3.1.2 Directories

Purpose. Here you will learn about the Unix directory tree, how to manipulate it and how to move around in it.

A unix file system is a tree of directories, where a directory is a container for files or more directories. We will display directories as follows:

```
/ ..... The root of the directory tree
├─ bin ..... Binary programs
└─ home ..... Location of users directories
```

The root of the Unix directory tree is indicated with a slash. Do `ls /` to see what the files and directories there are in the root. Note that the root is not the location where you start when you reboot your personal machine, or when you log in to a server.

Exercise. The command to find out your current working directory is `pwd`. Your home directory is your working directory immediately when you log in. Find out your home directory.

Expected outcome. You will typically see something like `/home/yourname` or `/Users/yourname`. This is system dependent.

Do `ls` to see the contents of the working directory. In the displays in this section, directory names will be followed by a slash: `dir/`[2], but this character is not part of their name. Example:

```
/home/you/
├── adirectory/
└── afile
```

The command for making a new directory is `mkdir`.

Exercise. Make a new directory with `mkdir somedir` and view the current directory with `ls`

Expected outcome. You should see this:

```
/home/you/
└── newdir/ ..... the new directory
```

The command for going into another directory, that is, making it your working directory, is `cd` ('change directory'). It can be used in the following ways:

- `cd` Without any arguments, `cd` takes you to your home directory.
- `cd <absolute path>` An absolute path starts at the root of the directory tree, that is, starts with `/`. The `cd` command takes you to that location.
- `cd <relative path>` A relative path is one that does not start at the root. This form of the `cd` command takes you to `<yourcurrentdir>/<relative path>`.

Exercise. Do `cd newdir` and find out where you are in the directory tree with `pwd`. Confirm with `ls` that the directory is empty. How would you get to this location using an absolute path?

Expected outcome. `pwd` should tell you `/home/you/newdir`, and `ls` then has no output, meaning there is nothing to list. The absolute path is `/home/you/newdir`.

Exercise. Let's quickly create a file in this directory: `touch onefile`, and another directory: `mkdir otherdir`. Do `ls` and confirm that there are a new file and directory.

2. You can tell your shell to operate like this by stating `alias ls=ls -F` at the start of your session.

Expected outcome. You should now have:

```
/home/you/
└── newdir/.....................................................you are here
    ├── onefile
    └── otherdir/
```

The `ls` command has a very useful option: with `ls -a` you see your regular files and hidden files, which have a name that starts with a dot. Doing `ls -a` in your new directory should tell you that there are the following files:

```
/home/you/
└── newdir/.....................................................you are here
    ├── .
    ├── ..
    ├── onefile
    └── otherdir/
```

The single dot is the current directory, and the double dot is the directory one level back.

Exercise. Predict where you will be after `cd ./otherdir/..` and check to see if you were right.

Expected outcome. The single dot sends you to the current directory, so that does not change anything. The `otherdir` part makes that subdirectory your current working directory. Finally, `..` goes one level back. In other words, this command puts your right back where you started.

Since your home directory is a special place, there are shortcuts for `cd`'ing to it: `cd` without arguments, `cd ~`, and `cd $HOME` all get you back to your home.

Go to your home directory, and from there do `ls newdir` to check the contents of the first directory you created, without having to go there.

Exercise. What does `ls ..` do?

Expected outcome. Recall that `..` denotes the directory one level up in the tree: you should see your own home directory, plus the directories of any other users.

Exercise. Can you use `ls` to see the contents of someone else's home directory? In the previous exercise you saw whether other users exist on your system. If so, do `ls ../thatotheruser`.

Expected outcome. If this is your private computer, you can probably view the contents of the other user's directory. If this is a university computer or so, the other directory may very well be protected – permissions are discussed in the next section – and you get `ls: ../otheruser: Permission denied`.

Make an attempt to move into someone else's home directory with `cd`. Does it work?

Victor Eijkhout

You can make copies of a directory with `cp`, but you need to add a flag to indicate that you recursively copy the contents: `cp -r`. Make another directory `somedir` in your home so that you have

```
/home/you/
├── newdir/ ..... you have been working in this one
└── somedir/ ..... you just created this one
```

What is the difference between `cp -r newdir somedir` and `cp -r newdir thirddir` where `thirddir` is not an existing directory name?

B.3.1.3 Permissions

Purpose. In this section you will learn about how to give various users on your system permission to do (or not to do) various things with your files.

Unix files, including directories, have permissions, indicating 'who can do what with this file'. Actions that can be performed on a file fall into three categories:

- reading `r`: any access to a file (displaying, getting information on it) that does not change the file;
- writing `w`: access to a file that changes its content, or even its metadata such as 'date modified';
- executing `x`: if the file is executable, to run it; if it is a directory, to enter it.

The people who can potentially access a file are divided into three classes too:

- the user `u`: the person owning the file;
- the group `g`: a group of users to which the owner belongs;
- other `o`: everyone else.

These nine permissions are rendered in sequence

user	*group*	*other*
rwx	*rwx*	*rwx*

For instance `rw-r--r--` means that the owner can read and write a file, the owner's group and everyone else can only read. The third letter, `w`, has two meanings: for files it indicates that they can be executed, for directories it indicates that they can be entered.

Permissions are also rendered numerically in groups of three bits, by letting `r` $= 4$, `w` $= 2$, `x` $= 1$:

rwx
421

Common codes are $7 =$ `rwx` and $6 =$ `rw`. You will find many files that have permissions 755 which stands for an executable that everyone can run, but only the owner can change, or 644 which stands for a data file that everyone can see but again only the owner can alter. You can set permissions by

```
chmod <permissions> file          # just one file
chmod -R <permissions> directory # directory, recursively
```

Examples:

```
chmod 755 file  # set to rwxrw-rw-
chmod g+w file  # give group write permission
chmod g=rx file # set group permissions
chod o-w  file  # take away write permission from others
chmod o=  file  # take away all permissions from others.
```

The man page gives all options.

Exercise. Make a file `foo` and do `chmod u-r foo`. Can you now inspect its contents? Make the file readable again, this time using a numeric code. Now make the file readable to your classmates. Check by having one of them read the contents.

Expected outcome. When you've made the file 'unreadable' by yourself, you can still `ls` it, but not `cat` it: that will give a 'permission denied' message.

Make a file `com` with the following contents:

```
#!/bin/sh
echo "Hello world!"
```

This is a legitimate shell script. What happens when you type `./com`? Can you make the script executable?

Adding or taking away permissions can be done with the following syntax:

```
chmod g+r,o-x file # give group read permission
                   # remove other execute permission
```

B.3.1.4 Wildcards

You already saw that `ls filename` gives you information about that one file, and `ls` gives you all files in the current directory. To see files with certain conditions on their names, the *wildcard* mechanism exists. The following wildcards exist:

- `*` any number of characters.
- `?` any character.

Example:

```
%% ls
s       sk      ski     skiing  skill
%% ls ski*
ski     skiing  skill
```

B.3.2 Text searching and regular expressions

Purpose. In this section you will learn how to search for text in files.

For this section you need at least one file that contains some amount of text. You can for instance get random text from `http://www.lipsum.com/feed/html`.

The `grep` command can be used to search for a text expression in a file.

Exercise. Search for the letter `q` in your text file with `grep q yourfile` and search for it in all files in your directory with `grep q *`. Try some other searches.

Expected outcome. In the first case, you get a listing of all lines that contain a `q`; in the second case, `grep` also reports what file name the match was found in: `qfile:this line has q in it`.

Caveats. If the string you are looking for does not occur, `grep` will simply not output anything. Remember that this is standard behaviour for Unix commands if there is nothing to report.

In addition to searching for literal strings, you can look for more general expressions.

`^`	the beginning of the line
`$`	the end of the line
`.`	any character
`*`	any number of repetitions
`[xyz]`	any of the characters `xyz`

This looks like the wildcard mechanism you just saw (section B.3.1.4) but it's subtly different. Compare the example above with:

```
%% cat s
sk
ski
skill
skiing
%% grep "ski*" s
sk
ski
skill
skiing
```

Some more examples: you can find

- All lines that contain the letter 'q' with `grep q yourfile`;
- All lines that start with an 'a' with `grep "^a" yourfile` (if your search string contains special characters, it is a good idea to use quote marks to enclose it);
- All lines that end with a digit with `grep "[0-9]$" yourfile`.

Exercise. Construct the search strings for finding

- lines that start with an uppercase character, and
- lines that contain exactly one character.

Expected outcome. For the first, use the range characters `[]`, for the second use the period to match any character.

The star character stands for zero or more repetitions of the character it follows, so that `a*` stands for any, possibly empty, string of `a` characters.

Exercise. Add a few lines `x = 1`, `x  = 2`, `x   = 3` (that is, have different numbers of spaces between `x` and the equals sign) to your test file, and make `grep` commands to search for all assignments to `x`.

The characters in the table above have special meanings. If you want to search that actual character, you have to *escape* it.

Exercise. Make a test file that has both `abc` and `a.c` in it, on separate lines. Try the commands `grep "a.c" file`, `grep a\.c file`, `grep "a\.c" file`.

Expected outcome. You will see that the period needs to be escaped, and the search string needs to be quoted. In the absence of either, you will see that `grep` also finds the `abc` string.

B.3.2.1 The system and other users

If you are on your personal machine, you may be the only user logged in. On university machines or other servers, there will often be other users. Here are some commands relating to them.

`whoami` show your login name
`who` show the other currently logged in users
`finger otherusername` get information about another user
`top` which processes are running on the system; use `top -u` to get this sorted the amount of cpu time they are currently taking. (On Linux, try also the `vmstat` command.)
`uptime` how long has your system been up?

B.3.3 Command execution

B.3.3.1 Search paths

Purpose. In this section you will learn how Unix determines what to do when you type a command name.

If you type a command such as `ls`, the shell does not just rely on a list of commands: it will actually go searching for a program by the name `ls`. This means that you can have multiple different commands with the same name, and which one gets executed depends on which one is found first.

Exercise. What you may think of as 'Unix commands' are often just executable files in a system directory. Do `which cd`, and do an `ls -l` on the result

Expected outcome. The location of `cd` is something like `/usr/bin/cd`. If you `ls` that, you will see that it is probably owned by root. Its executable bits are probably set for all users.

The locations where unix searches for commands is the 'search path'

Exercise. Do `echo $PATH`. Can you find the location of `cd`? Are there other commands in the same location? Is the current directory '`.`' in the path? If not, do `export PATH=".:$PATH"`. Now create an executable file `cd` in the current director (see above for the basics), and do `cd`. Some people consider having the working directory in the path a security risk.

Expected outcome. The path will be a list of colon-separated directories,
for instance `/usr/bin:/usr/local/bin:/usr/X11R6/bin`. If the working directory is in the path, it will probably be at the end: `/usr/X11R6/bin:.` but most likely it will not be there. If you put '`.`' at the start of the path, unix will find the local `cd` command before the system one.

It is possible to define your own commands as aliases of existing commands.

Exercise. Do `alias chdir=cd` and convince yourself that now `chdir` works just like `cd`. Do `alias rm='rm -i'`; look up the meaning of this in the man pages. Some people find this alias a good idea; can you see why?

Expected outcome. The `-i` 'interactive' option for `rm` makes the command ask for confirmation before each delete. Since unix does not have a trashcan that needs to be emptied explicitly, this can be a good idea.

B.3.3.2 Redirection and Pipelines

Purpose. In this section you will learn how to feed one command into another, and how to connect commands to input and output files.

So far, the unix commands you have used have taken their input from your keyboard, or from a file named on the command line; their output went to your screen. There are other possibilities for providing input from a file, or for storing the output in a file.

B.3.3.2.1 **Input redirection** The `grep` command had two arguments, the second being a file name. You can also write `grep string < yourfile`, where the less-than sign means that the input will come from the named file, `yourfile`.

B.3.3.2.2 **Output redirection** More usefully, `grep string yourfile > outfile` will take what normally goes to the terminal, and send it to `outfile`. The output file is created if it didn't already exist, otherwise it is overwritten. (To append, use `grep text yourfile >> outfile`.)

Exercise. Take one of the grep commands from the previous section, and send its output to a file. Check that the contents of the file are identical to what appeared on your screen before. Search for a string that does not appear in the file and send the output to a file. What does this mean for the output file?

Expected outcome. Searching for a string that does not occur in a file gives no terminal output. If you redirect the output of this `grep` to a file, it gives a zero size file. Check this with `ls` and `wc`.

B.3.3.2.3 **Standard files** Unix has three standard files that handle input and output:

`stdin` is the file that provides input for processes.
`stdout` is the file where the output of a process is written.
`stderr` is the file where error output is written.

In an interactive session, all three files are connected to the user terminal. Using inpur or output redirection then means that the input is taken or the output sent to a different file than the terminal.

B.3.3.2.4 **Command redirection** Instead of taking input from a file, or sending output to a file, it is possible to connect two commands together, so that the second takes the output of the first as input. The syntax for this is `cmdone | cmdtwo`; this is called a pipeline. For instance, `grep a yourfile | grep b` finds all lines that contains both an `a` and a `b`.

Exercise. Construct a pipeline that counts how many lines there are in your file that contain the string `th`. Use the `wc` command (see above) to do the counting.

There are a few more ways to combine commands. Suppose you want to present the result of `wc` a bit nicely. Type the following command

```
echo The line count is wc -l foo
```

where `foo` is the name of an existing file. The way to get the actual line count echoed is by the backquote:

```
echo The line count is `wc -l foo`
```

Anything in between backquotes is executed before the rest of the command line is evaluated. The way `wc` is used here, it prints the file name. Can you find a way to prevent that from happening?

B.3.3.3 Processes

The Unix operating system can run many programs at the same time, by rotating through the list and giving each only a fraction of a second to run each time. The command `ps` can tell you everything that is currently running.

Exercise. Type `ps`. How many programs are currently running? By default `ps` gives you only programs that you explicitly started. Do `ps guwax` for a detailed list of everything that is running. How many programs are running? How many belong to the root user, how many to you?

Expected outcome. To count the programs belonging to a user, pipe the `ps` command through an appropriate `grep`, which can then be piped to `wc`.

In this long listing of `ps`, the second column contains the process numbers. Sometimes it is useful to have those. The `cut` command can cut certain position from a line: type `ps guwax | cut -c 10-14`.

To get dynamic information about all running processes, use the `top` command. Read the man page to find out how to sort the output by CPU usage.

When you type a command and hit return, that command becomes, for the duration of its run, the *foreground process*. Everything else that is running at the same time is a *background process*.

Make an executable file `hello` with the following contents:

```
#!/bin/sh
while [ 1 ] ; do
  sleep 2
  date
done
```

and type `./hello`.

Exercise. Type `Control-z`. This suspends the foreground process. It will give you a number like `[1]` or `[2]` indicating that it is the first or second program that has been suspended or put in the background. Now type `bg` to put this process in the background. Confirm that there is no foreground process by hitting return, and doing an `ls`.

Expected outcome. After you put a process in the background, the terminal is available again to accept foreground commands. If you hit return, you should see the command prompt.

Exercise. Type `jobs` to see the processes in the current session. If the process you just put in the background was number 1, type `fg %1`. Confirm that it is a foreground process again.

Expected outcome. If a shell is executing a program in the foreground, it will not accept command input, so hitting return should only produce blank lines.

Exercise. When you have made the `hello` script a foreground process again, you can kill it with `Control-c`. Try this. Start the script up again, this time as `./hello &` which immediately puts it in the background. You should also get output along the lines of `[1] 12345` which tells you that it is the first job you put in the background, and that `12345` is its process ID. Kill the script with `kill %1`. Start it up again, and kill it by using the process number.

Expected outcome. The command `kill 12345` using the process number is usually enough to kill a running program. Sometimes it is necessary to use `kill -9 12345`.

B.3.3.4 Shell customization

Above it was mentioned that `ls -F` is an easy way to see which files are regular, executable, or directories; by typing `alias ls='ls -F'` the `ls` command will automatically expanded to `ls -F` every time it is invoked. If you would like this behaviour in every login session, you can add the `alias` command to your `.profile` file. Other shells have other files for such customizations.

B.3.4 Scripting

The unix shells are also programming environments. You will learn more about this aspect of unix in this section.

B.3.4.1 Shell variables

Above you encountered `PATH`, which is an example of an shell, or environment, variable. These are variables that are known to the shell and that can be used by all programs run by the shell. You can see the full list of all variables known to the shell by typing `env`.

You can get the value of a shell variable by prefixing it with a dollar sign. Type the following two commands and compare the output:

```
echo PATH
echo $PATH
```

Exercise. Check on the value of the `HOME` variable by typing `echo $HOME`. Also find the value of `HOME` by piping `env` through `grep`.

B.3.4.2 Control structures

Like any good programming system, the shell has some control structures. Their syntax takes a bit of getting used to. (Different shells have different syntax; in this tutorial we only discuss the bash shell.

In the bash shell, control structures can be written over several lines:

```
if [ $PATH = "" ] ; then
  echo "Error: path is empty"
fi
```

or on a single line:

```
if [ `wc -l file` -gt 100 ] ; then echo "file too long" ; fi
```

There are a number of tests defined, for instance `-f somefile` tests for the existence of a file. Change your script so that it will report `-1` if the file does not exist.

There are also loops. A `for` loop looks like

```
for var in listofitems ; do
  something with $var
done
```

This does the following:

- for each item in `listofitems`, the variable `var` is set to the item, and
- the loop body is executed.

As a simple example:

```
   %% for x in a b c ; do echo $x ; done
a
b
c
```

In a more meaningful example, here is how you would make backups of all your `.c` files:

```
for cfile in *.c ; do
  cp $cfile $cfile.bak
done
```

Shell variables can be manipulated in a number of ways. Execute the following commands to see that you can remove trailing characters from a variable:

```
%% a=b.c
%% echo ${a%.c}
b
```

With this as a hint, write a loop that renames all your `.c` files to `.x` files.

B.3.4.3 Scripting

It is possible to write programs of unix shell commands. First you need to know how to put a program in a file and have it be executed. Make a file `script1` containing the following two lines:

```
#!/bin/bash
echo "hello world"
```

and type `./script1` on the command line. Result? Make the file executable and try again.

You can give your script command line arguments `./script1 foo bar`; these are available as `$1` et cetera in the script.

Write a script that takes as input a file name argument, and reports how many lines are in that file.

Edit your script to test whether the file has less than 10 lines (use the `foo -lt bar` test), and if it does, `cat` the file. Hint: you need to use backquotes inside the test.

The number of command line arguments is available as `$#`. Add a test to your script so that it will give a helpful message if you call it without any arguments.

B.3.5 Expansion

The shell performs various kinds of expansion on a command line, that is, replacing part of the commandline with different text.

Brace expansion:

```
%% echo a{b,cc,ddd}e
abe acce addde
```

This can for instance be used to delete all extension of some base file name:

```
%% rm tmp.{c,s,o}  # delete tmp.c tmp.s tmp.o
```

Tilde expansion gives your own, or someone else's home directory:

```
%% echo ~
/share/home/00434/eijkhout
%% echo ~eijkhout
/share/home/00434/eijkhout
```

Parameter expansion gives the value of shell variables:

```
%% x=5
%% echo $x
5
```

Undefined variables do not give an error message:

```
%% echo $y
```

There are many variations on parameter expansion. Above you already saw that you can strip trailing characters:

```
%% a=b.c
%% echo ${a%.c}
b
```

Here is how you can deal with undefined variables:

```
%% echo ${y:-0}
0
```

The backquote mechanism (section B.3.3.2 above) is known as command substitution:

```
%% echo 123 > w
%% cat w
123
%% wc -c w
       4 w
```

```
%% if [ `cat w | wc -c` -eq 4 ] ; then echo four ; fi
four
```

Unix shell programming is very much oriented towards text manipulation, but it is possible to do arithmetic. Arithmetic substitution tells the shell to treat the expansion of a parameter as a number:

```
%% x=1
%% echo $((x*2))
2
```

Integer ranges can be used as follows:

```
%% for i in {1..10} ; do echo $i ; done
1
2
3
4
5
6
7
8
9
10
```

B.3.6 Shell interaction

Interactive use of Unix, in contrast to script writing (section B.3.4), is a complicated conversation between the user and the shell. You, the user, type a line, hit return, and the shell tries to interpret it. There are several cases.

- Your line contains one full command, such as `ls foo`: the shell will execute this command.
- You can put more than one command on a line, separated by semicolons: `mkdir foo; cd foo`. The shell will execute these commands in sequence.
- Your input line is not a full command, for instance `while [ 1]`. The shell will recognize that there is more to come, and use a different prompt to show you that it is waiting for the remainder of the command.
- Your input line would be a legitimate command, but you want to type more on a second line. In that case you can end your input line with a backslash character, and the shell will recognize that it needs to hold off on executing your command. In effect, the backslash will hide (*escape*) the return.

When the shell has collected a command line to execute, by using one or more of your input line or only part of one, as described just now, it will apply expansion to the command line (section B.3.5). It will then interpret the commandline as a command and arguments, and proceed to invoke that command with the arguments as found.

There are some subtleties here. If you type `ls *.c`, then the shell will reognize the wildcard character and expand it to a command line, for instance `ls foo.c bar.c`. Then it will invoke the `ls` command with the argument list `foo.c bar.c`. Note that `ls` does not receive `*.c` as argument! In cases where you do want the unix command to receive an argument with a wildcard, you need to escape it so that the shell will not expand it. For instance, `find . -name .c` will make the shell invoke `find` with arguments `. -name *.c`.

B.3.7 Review questions

Exercise 2.1. Write a shell script for making backups. When you call is with `./backup somefile` it should test whether `somefile.bak` exists, and if not, copy the one to the other. (This is crude, if the backup file already exists, it is overwritten.)

B.4 Compilers and libraries

B.4.1 An introduction to binary files

Purpose. In this section you will be introduced to the different types of binary files that you encounter while programming.

One of the first things you become aware of when you start programming is the distinction between the readable source code, and the unreadable, but executable, program code. In this tutorial you will learn about a couple more file types:

- A source file can be compiled to an *object file*, which is a bit like a piece of an executable: by itself it does nothing, but it can be combined with other object files to form an executable.
- A *library* is a bundle of object files that can be used to form an executable. Often, libraries are written by an expert and contain code for specialized purposes such as linear algebra manipulations. Libraries are important enough that they can be commercial, to be bought if you need expert code for a certain purpose.

You will now learn how these types of files are created and used.

B.4.2 Simple compilation

Purpose. In this section you will learn about executables and object files.

Let's start with a simple program that has the whole source in one file.

One file: `hello.c`

```
#include <stdlib.h>
#include <stdio.h>

int main() {
  printf("hello world\n");
  return 0;
}
```

Compile this program with your favourite compiler; we will use `gcc` in this tutorial, but substitute your own as desired. As a result of the compilation, a file `a.out` is created, which is the executable.

```
%% gcc hello.c
%% ./a.out
hello world
```

You can get a more sensible program name with the `-o` option:

```
%% gcc -o helloprog hello.c
%% ./helloprog
hello world
```

Now we move on to a program that is in more than one source file.

Main program: `fooprog.c`

```
#include <stdlib.h>
#include <stdio.h>

extern bar(char*);

int main() {
  bar("hello world\n");
  return 0;
}
```

Subprogram: `fooprog.c`

```
#include <stdlib.h>
#include <stdio.h>

void bar(char *s) {
  printf("%s",s);
  return;
}
```

As before, you can make the program with one command.

```
%% gcc -o foo fooprog.c foosub.c
%% ./foo
hello world
```

However, you can also do it in steps, compiling each file separately and then linking them together.

```
%% gcc -c fooprog.c
%% gcc -c foosub.c
%% gcc -o foo fooprog.o foosub.o
%% ./foo
hello world
```

The `-c` option tells the compiler to compile the source file, giving an *object file*. The third command than acts as the *linker*, tieing together the object files into an executable. (With programs that are spread over several files there is always the danger of editing a subroutine definition and then forgetting to update all the places it is used. See the 'make' tutorial, section B.5, for a way of dealing with this.)

B.4.3 Libraries

Purpose. In this section you will learn about libraries.

If you have written some subprograms, and you want to share them with other people (perhaps by selling them), then handing over individual object files is inconvenient. Instead, the solution is to combine them into a library. First

we look at *static libraries*, for which the *archive utility* `ar` is used. A static library is linked into your executable, becoming part of it. This may lead to large executables; you will learn about shared libraries next, which do not suffer from this problem.

Create a directory to contain your library (depending on what your library is for this can be a system directory such as `/usr/bin`), and create the library file there.

```
%% mkdir ../lib
%% ar cr ../lib/libfoo.a foosub.o
```

The `nm` command tells you what's in the library:

```
%% nm ../lib/libfoo.a

../lib/libfoo.a(foosub.o):
00000000 T _bar
         U _printf
```

Line with `T` indicate functions defined in the library file; a `U` indicates a function that is used.

The library can be linked into your executable by explicitly giving its name, or by specifying a library path:

```
%% gcc -o foo fooprog.o ../lib/libfoo.a
# or
%% gcc -o foo fooprog.o -L../lib -lfoo
%% ./foo
hello world
```

A third possibility is to use the `LD_LIBRARY_PATH` shell variable. Read the man page of your compiler for its use, and give the commandlines that create the `foo` executable, linking the library through this path.

Although they are somewhat more complicated to use, *shared libraries* have several advantages. For instance, since they are not linked into the executable but only loaded at runtime, they lead to (much) smaller executables. They are not created with `ar`, but through the compiler. For instance:

```
%% gcc -dynamiclib -o ../lib/libfoo.so foosub.o
%% nm ../lib/libfoo.so

../lib/libfoo.so(single module):
00000fc4 t __dyld_func_lookup
00000000 t __mh_dylib_header
00000fd2 T _bar
         U _printf
00001000 d dyld__mach_header
00000fb0 t dyld_stub_binding_helper
```

Shared libraries are not actually linked into the executable; instead, the executable will contain the information where the library is to be found at execution time:

```
%% gcc -o foo fooprog.o -L../lib -Wl,-rpath,`pwd`/../lib -lfoo
%% ./foo
hello world
```

The link line now contains the library path twice:

1. once with the `-L` directive so that the linker can resolve all references, and
2. once with the linker directive `-Wl,-rpath,`pwd`/../lib` which stores the path into the executable so that it can be found at runtime.

Build the executable again, but without the `-Wl` directive. Where do things go wrong and why?

Use the command `ldd` to get information about what shared libraries your executable uses. (On Mac OS X, use `otool -L` instead.)

B.5 Managing programs with Make

The *Make* utility helps you manage the building of projects: its main task is to facilitate rebuilding only those parts of a multi-file project that need to be recompiled or rebuilt. This can save lots of time, since it can replace a minutes-long full installation by a single file compilation. *Make* can also help maintaining multiple installations of a program on a single machine, for instance compiling a library with more than one compiler, or compiling a program in debug and optimized mode.

Make is a Unix utility with a long history, and traditionally there are variants with slightly different behaviour, for instance on the various flavours of Unix such as HP-UX, AUX, IRIX. These days, it is advisable, no matter the platform, to use the GNU version of Make which has some very powerful extensions; it is available on all Unix platforms (on Linux it is the only available variant), and it is a *de facto* standard. The manual is available at `http://www.gnu.org/software/make/manual/make.html`, or you can read the book [63].

There are other build systems, most notably Scons and Bjam. We will not discuss those here. The examples in this tutorial will be for the C and Fortran languages, but *Make* can work with any language, and in fact with things like TEX that are not really a language at all; see section B.5.6.

B.5.1 A simple example

Purpose. In this section you will see a simple example, just to give the flavour of *Make*.

B.5.1.1 C

Make the following files:

foo.c

```
#include "bar.h"
int c=3;
int d=4;
int main()
{
  int a=2;
  return(bar(a*c*d));
}
```

bar.c

```
#include "bar.h"
int bar(int a)
{
  int b=10;
  return(b*a);
}
```

bar.h

```
extern int bar(int);
```

and a makefile:

Makefile

```
fooprog : foo.o bar.o
	cc -o fooprog foo.o bar.o
foo.o : foo.c
	cc -c foo.c
bar.o : bar.c
	cc -c bar.c
clean :
	rm -f *.o fooprog
```

The makefile has a number of rules like

```
foo.o : foo.c
<TAB>cc -c foo.c
```

which have the general form

```
target : prerequisite(s)
<TAB>rule(s)
```

where the rule lines are indented by a `TAB` character.

A rule, such as above, states that a 'target' file `foo.o` is made from a 'prerequisite' `foo.c`, namely by executing the command `cc -c foo.c`. The precise definition of the rule is:

- if the target `foo.o` does not exist or is older than the prerequisite `foo.c`,
- then the command part of the rule is executed: `cc -c foo.c`
- If the prerequisite is itself the target of another rule, than that rule is executed first.

Probably the best way to interpret a rule is:

- if any prerequisite has changed,
- then the target needs to be remade,
- and that is done by executing the commands of the rule.

If you call `make` without any arguments, the first rule in the makefile is evaluated. You can execute other rules by explicitly invoking them, for instance `make foo.o` to compile a single file.

Exercise. Call `make`.

Expected outcome. The above rules are applied: `make` without arguments tries to build the first target, `fooprog`. In order to build this, it needs the prerequisites `foo.o` and `bar.o`, which do not exist. However, there are rules for making them, which `make` recursively invokes. Hence you see two compilations, for `foo.o` and `bar.o`, and a link command for `fooprog`.

Caveats. Typos in the makefile or in file names can cause various errors. In particular, make sure you use tabs and not spaces for the rule lines. Unfortunately, debugging a makefile is not simple. *Make*'s error message will usually give you the line number in the make file where the error was detected.

Exercise. Do `make clean`, followed by `mv foo.c boo.c` and `make` again. Explain the error message. Restore the original file name.

Expected outcome. *Make* will complain that there is no rule to make `foo.c`. This error was caused when `foo.c` was a prerequisite, and was found not to exist. *Make* then went looking for a rule to make it.

Now add a second argument to the function `bar`. This requires you to edit `bar.c` and `bar.h`: go ahead and make these edits. However, it also requires you to edit `foo.c`, but let us for now 'forget' to do that. We will see how *Make* can help you find the resulting error.

Exercise. Call `make` to recompile your program. Did it recompile `foo.c`?

Expected outcome. Even through conceptually `foo.c` would need to be recompiled since it uses the `bar` function, *Make* did not do so because the makefile had no rule that forced it.

In the makefile, change the line

```
foo.o : foo.c
```

to

```
foo.o : foo.c bar.h
```

which adds `bar.h` as a prerequisite for `foo.o`. This means that, in this case where `foo.o` already exists, *Make* will check that `foo.o` is not older than any of its prerequisites. Since `bar.h` has been edited, it is younger than `foo.o`, so `foo.o` needs to be reconstructed.

Exercise. Confirm that the new makefile indeed causes `foo.o` to be recompiled if `bar.h` is changed. This compilation will now give an error, since you 'forgot' to edit the use of the `bar` function.

B.5.1.2 Fortran

Make the following files:

foomain.F

```
      program test
      use testmod

      call func(1,2)

      end program
```

foomod.F

```
      module testmod

      contains

      subroutine func(a,b)
      integer a,b
      print *,a,b,c
      end subroutine func

      end module
```

and a makefile:

Makefile

```
fooprog : foomain.o foomod.o
       gfortran -o fooprog foo.o foomod.o
foomain.o : foomain.F
       gfortran -c foomain.F
foomod.o : foomod.F
       gfortran -c foomod.F
clean :
       rm -f *.o fooprog
```

If you call `make`, the first rule in the makefile is executed. Do this, and explain what happens.

Exercise. Call `make`.

Expected outcome. The above rules are applied: `make` without arguments tries to build the first target, `foomain`. In order to build this, it needs the prerequisites `foomain.o` and `foomod.o`, which do not exist. However, there are rules for making them, which `make` recursively invokes. Hence you see two compilations, for `foomain.o` and `foomod.o`, and a link command for `fooprog`.

Caveats. Typos in the makefile or in file names can cause various errors. Unfortunately, debugging a makefile is not simple. You will just have to understand the errors, and make the corrections.

Exercise. Do `make clean`, followed by `mv foo.c boo.c` and `make` again. Explain the error message. Restore the original file name.

Expected outcome. Make will complain that there is no rule to make `foo.c`. This error was caused when `foo.c` was a prerequisite, and was found not to exist. *Make* then went looking for a rule to make it.

Now add an extra parameter to `func` in `foomod.F` and recompile.

Exercise. Call `make` to recompile your program. Did it recompile `foomain.F`?

Expected outcome. Even through conceptually `foomain.F` would need to be recompiled, *Make* did not do so because the makefile had no rule that forced it.

Change the line

```
foomain.o : foomain.F
```

to

```
foomain.o : foomain.F foomod.F
```

which adds `foomod.F` as a prerequisite for `foomain.o`. This means that, in this case where `foomain.o` already exists, *Make* will check that `foomain.o` is not older than any of its prerequisites. Since `foomod.F` has ben edited, it is younger than `foomain.o`, so `foomain.o` needs to be reconstructed.

Exercise. Confirm that the corrected makefile indeed causes `foomain.F` to be recompiled.

B.5.2 Variables and template rules

Purpose. In this section you will learn various work-saving mechanism in *Make*, such as the use of variables, and of template rules.

B.5.2.1 Makefile variables

It is convenient to introduce variables in your makefile. For instance, instead of spelling out the compiler explicitly every time, introduce a variable in the makefile:

```
CC = gcc
FC = gfortran
```

and use `${CC}` or `${FC}` on the compile lines:

```
foo.o : foo.c
        ${CC} -c foo.c
foomain.o : foomain.F
        ${FC} -c foomain.F
```

Exercise. Edit your makefile as indicated. First do `make clean`, then `make foo` (C) or `make fooprog` (Fortran).

Expected outcome. You should see the exact same compile and link lines as before.

Caveats. Unlike in the shell, where braces are optional, variable names in a makefile have to be in braces or parentheses. Experiment with what hapens if you forget the braces around a variable name.

One advantage of using variables is that you can now change the compiler from the commandline:

```
make CC="icc -O2"
make FC="gfortran -g"
```

Exercise. Invoke *Make* as suggested (after `make clean`). Do you see the difference in your screen output?

Expected outcome. The compile lines now show the added compiler option `-O2` or `-g`.

Make also has built-in variables:

- `$@` The target. Use this in the link line for the main program.
- `$^` The list of prerequisites. Use this also in the link line for the program.
- `$<` The first prerequisite. Use this in the compile commands for the individual object files.

Using these variables, the rule for `fooprog` becomes

```
fooprog : foo.o bar.o
        ${CC} -o $@ $^
```

and a typical compile line becomes

```
foo.o : foo.c bar.h
        ${CC} -c $<
```

You can also declare a variable

```
THEPROGRAM = fooprog
```

and use this variable instead of the program name in your makefile. This makes it easier to change your mind about the name of the executable later.

Exercise. Construct a commandline so that your makefile will build the executable `fooprog_v2`.

Expected outcome. `make THEPROGRAM=fooprog_v2`

B.5.2.2 Template rules

In your makefile, the rules for the object files are practically identical:

- the rule header (`foo.o : foo.c`) states that a source file is a prerequisite for the object file with the same base name;
- and the instructions for compiling (`${CC} -c $<`) are even character-for-character the same, now that you are using *Make*'s built-in variables;
- the only rule with a difference is
  ```
  foo.o : foo.c bar.h
          ${CC} -c $<
  ```
  where the object file depends on the source file and another file.

We can take the commonalities and summarize them in one rule[3]:

3. This mechanism is the first instance you'll see that only exists in GNU make, though in this particular case there is a similar mechanism in standard make. That will not be the case for the wildcard mechanism in the next section.

```
%.o : %.c
        ${CC} -c $<
%.o : %.F
        ${FC} -c $<
```

This states that any object file depends on the C or Fortran file with the same base name. To regenerate the object file, invoke the C or Fortran compiler with the `-c` flag. These template rules can function as a replacement for the multiple specific targets in the makefiles above, except for the rule for `foo.o`.

The dependence of `foo.o` on `bar.h` can be handled by adding a rule

```
foo.o : bar.h
```

with no further instructions. This rule states, 'if file `bar.h` changed, file `foo.o` needs updating'. *Make* will then search the makefile for a different rule that states how this updating is done.

Exercise. Change your makefile to incorporate these ideas, and test.

B.5.3 Wildcards

Your makefile now uses one general rule for compiling all your source files. Often, these source files will be all the `.c` or `.F` files in your directory, so is there a way to state 'compile everything in this directory'? Indeed there is. Add the following lines to your makefile, and use the variable `COBJECTS` or `FOBJECTS` wherever appropriate.

```
# wildcard: find all files that match a pattern
CSOURCES := ${wildcard *.c}
# pattern substitution: replace one pattern string by another
COBJECTS := ${patsubst %.c,%.o,${SRC}}

FSOURCES := ${wildcard *.F}
FOBJECTS := ${patsubst %.F,%.o,${SRC}}
```

B.5.4 Miscellania

B.5.4.1 What does this makefile do?

Above you learned that issuing the `make` command will automatically execute the first rule in the makefile. This is convenient in one sense[4], and inconvenient in another: the only way to find out what possible actions a makefile allows is to read the makefile itself, or the – usually insufficient – documentation.

A better idea is to start the makefile with a target

4. There is a convention among software developers that a package can be installed by the sequence `./configure ; make ; make install`, meaning: Configure the build process for this computer, Do the actual build, Copy files to some system directory such as `/usr/bin`.

```
info :
        @echo "The following are possible:"
        @echo "  make"
        @echo "  make clean"
```

Now `make` without explicit targets informs you of the capabilities of the makefile.

B.5.4.2 Phony targets

The example makefile contained a target `clean`. This uses the *Make* mechanisms to accomplish some actions that are not related to file creation: calling `make clean` causes *Make* to reason 'there is no file called `clean`, so the following instructions need to be performed'. However, this does not actually cause a file `clean` to spring into being, so calling `make clean` again will make the same instructions being executed.

To indicate that this rule does not actually make the target, declare

```
.PHONY : clean
```

One benefit of declaring a target to be phony, is that the *Make* rule will still work, even if you have a file named `clean`.

B.5.4.3 Predefined variables and rules

Calling `make -p yourtarget` causes make to print out all its actions, as well as the values of all variables and rules, both in your makefile and ones that are predefined. If you do this in a directory where there is no makefile, you'll see that make actually already knows how to compile `.c` or `.F` files. Find this rule and find the definition of the variables in it.

You see that you can customize make by setting such variables as `CFLAGS` or `FFLAGS`. Confirm this with some experimentation. If you want to make a second makefile for the same sources, you can call `make -f othermakefile` to use this instead of the default `Makefile`.

Note, by the way, that both `makefile` and `Makefile` are legitimate names for the default makefile. It is not a good idea to have both `makefile` and `Makefile` in your directory.

B.5.4.4 Using the target as prerequisite

Suppose you have two different targets that are treated largely the same. You would want to write:

```
PROGS = myfoo other
${PROGS} : $@.o
        ${CC} -o $@ $@.o ${list of libraries goes here}
```

and saying `make myfoo` would cause

```
cc -c myfoo.c
cc -o myfoo myfoo.o ${list of libraries}
```

and likewise for `make other`. What goes wrong here is the use of `$@.o` as prerequisite. In Gnu Make, you can repair this as follows:

```
.SECONDEXPANSION:
${PROGS} : $$@.o
```

B.5.5 Shell scripting in a Makefile

Purpose. In this section you will see an example of a longer shell script appearing in a makefile rule.

In the makefiles you have seen so far, the command part was a single line. You can actually have as many lines there as you want. For example, let us make a rule for making backups of the program you are building.

Add a `backup` rule to your makefile. The first thing it needs to do is make a backup directory:

```
.PHONY : backup
backup :
        if [ ! -d backup ] ; then
          mkdir backup
        fi
```

Did you type this? Unfortunately it does not work: every line in the command part of a makefile rule gets executed as a single program. Therefore, you need to write the whole command on one line:

```
backup :
        if [ ! -d backup ] ; then mkdir backup ; fi
```

or if the line gets too long:

```
backup :
        if [ ! -d backup ] ; then \
          mkdir backup ; \
        fi
```

Next we do the actual copy:

```
backup :
        if [ ! -d backup ] ; then mkdir backup ; fi
        cp myprog backup/myprog
```

But this backup scheme only saves one version. Let us make a version that has the date in the name of the saved program.

The Unix `date` command can customize its output by accepting a format string. Type the following: `date` This can be used in the makefile.

Exercise. Edit the `cp` command line so that the name of the backup file includes the current date.

Expected outcome. Hint: you need the backquote. Consult the Unix tutorial if you do not remember what backquotes do.

If you are defining shell variables in the command section of a makefile rule, you need to be aware of the following. Extend your `backup` rule with a loop to copy the object files:

```
backup :
        if [ ! -d backup ] ; then mkdir backup ; fi
        cp myprog backup/myprog
        for f in ${OBJS} ; do \
          cp $f backup ; \
        done
```

(This is not the best way to copy, but we use it for the purpose of demonstration.) This leads to an error message, caused by the fact that *Make* interprets `$f` as an environment variable of the outer process. What works is:

```
backup :
        if [ ! -d backup ] ; then mkdir backup ; fi
        cp myprog backup/myprog
        for f in ${OBJS} ; do \
          cp $$f backup ; \
        done
```

(In this case *Make* replaces the double dollar by a single one when it scans the commandline. During the execution of the commandline, `$f` then expands to the proper filename.)

B.5.6 A Makefile for LaTeX

The *Make* utility is typically used for compiling programs, but other uses are possible too. In this section, we will discuss a makefile for LaTeX documents.

We start with a very basic makefile:

```
info :
@echo "Usage: make foo"
@echo "where foo.tex is a LaTeX input file"

%.pdf : %.tex
pdflatex $<
```

The command `make myfile.pdf` will invoke `pdflatex myfile.tex`, if needed, once. Next we repeat invoking `pdflatex` until the log file no longer reports that further runs are needed:

```
%.pdf : %.tex
pdflatex $<
while [ `cat ${basename $@}.log | grep "Rerun to get cross-references right." | wc
```

```
  pdflatex $< ; \
done
```

We use the `${basename fn}` macro to extract the base name without extension from the target name.

In case the document has a bibliography or index, we run `bibtex` and `makeindex`.

```
%.pdf : %.tex
pdflatex ${basename $@}
-bibtex ${basename $@}
-makeindex ${basename $@}
while [ `cat ${basename $@}.log | grep "Rerun to get cross-references right." | wc -l` -
  pdflatex ${basename $@} ; \
done
```

The minus sign at the start of the line means that *Make* should not abort if these commands fail.

Finally, we would like to use *Make*'s facility for taking dependencies into account. We could write a makefile that has the usual rules

```
mainfile.pdf : mainfile.tex includefile.tex
```

but we can also discover the include files explicitly. The following makefile is invoked with

```
  make pdf TEXTFILE=mainfile
```

The `pdf` rule then uses some shell scripting to discover the include files (but not recursively), and it calls emphMake again, invoking another rule, and passing the dependencies explicitly.

```
pdf :
export includes=`grep "^.input " ${TEXFILE}.tex | awk '{v=v FS $$2".tex"} END {print v}'
${MAKE} ${TEXFILE}.pdf INCLUDES="$$includes"

%.pdf : %.tex ${INCLUDES}
pdflatex $< ; \
while [ `cat ${basename $@}.log | grep "Rerun to get cross-references right." | wc -l` -
  pdflatex $< ; \
done
```

This shell scripting can also be done outside the makefile, generating the makefile dynamically.

B.6 Source control

Source code control systems, also called revision control systems, are a way of storing software, where not only the current version is stored, but also all previous versions. This is done by maintaining a *repository* for all versions, while one or more users work on a 'checked out' copy of the latest version. Those of the users that are developers can then commit their changes to the repository. Other users then update their local copy. The repository typically resides on a remote machine that is reliably backup up.

There are various reasons for keeping your source in a repository.

- If you work in a team, it is the best way to synchronize your work with your colleagues. It is also a way to document what changes were made, by whom, and why.
- It will allow you to roll back a defective code to a version that worked.
- It allows you to have branches, for instance for customizations that need to be kept out of the main development line. If you are working in a team, a branch is a way to develop a major feature, stay up to date with changes your colleagues make, and only add your feature to the main development when it is sufficiently tested.
- If you work alone, it is a way to synchronize between more than one machine. (You could even imagine traveling without all your files, and installing them from the repository onto a borrowed machine as the need arises.)
- Having a source code repository is one way to backup your work.

There are various source code control systems; in this tutorial you will learn the basics of *Subversion*, also called *svn*.

B.6.1 How to do this tutorial

This lab should be done two people, to simulate a group of programmers working on a joint project. You can also do this on your own by using two accounts on the same machine.

B.6.2 Create and populate a repository

Purpose. In this section you will create a repository and make a local copy to work on.

First we need to have a repository. In practice, you will often use one that has been set up by a sysadmin, but here one student will set it up in his home directory. Another option would be to use a hosting service, commercial or free such as `http://code.google.com/projecthosting`.

```
%% svnadmin create ./repository --fs-type=fsfs
```

For the purposes of this exercise we need to make this directory, and the home directory of the first student, visible to the other student. In a practical situation this is not necessary: the sysadmin would set up the repository on the server.

```
%% chmod -R g+rwX ./repository
%% chmod g+rX ~
```

Both students now make some dummy data to start the repository. Please choose unique directory names.

```
%%  mkdir sourcedir
%%  cat > sourcedir/firstfile
a
b
c
d
e
f
^D
```

This project data can now be added to the repository:

```
%%  svn import -m "initial upload" sourcedir file://`pwd`/repository
Adding         sourcedir/firstfile

Committed revision 1.
```

The second student needs to spell out the path

```
%% svn import -m "initial upload" otherdir \
    file:///share/home/12345/firststudent/repository
```

Instead of specifying the creation message on the commandline, you can also set the `EDITOR` or `SVN_EDITOR` environment variable and do

```
%%  svn import sourcedir file://`pwd`/repository
```

An editor will then open so that you can input descriptive notes.

Now we pretend that this repository is in some far location, and we want to make a local copy. Both students will now work on the same files in the repository, to simulate a programming team.

```
%% svn co file://`pwd`/repository myproject
A  myproject/secondfile
A  myproject/firstfile
Checked out revision 2.
```

(In practice, you would use `http:` or `svn+ssh:` as the protocol, rather than `file:`.) This is the last time you need to specify the repository path: the local copy you have made remembers the location.

Let's check that everything is in place:

```
%% ls
myproject/  repository/  sourcedir/
%% rm -rf sourcedir/
%% cd myproject/
%% ls -a
./  ../  firstfile  myfile  .svn/
```

```
%% svn info
Path: .
URL: file:///share/home/12345/yournamehere/repository
Repository UUID: 3317a5b0-6969-0410-9426-dbff766b663f
Revision: 2
Node Kind: directory
Schedule: normal
Last Changed Author: build
Last Changed Rev: 2
Last Changed Date: 2009-05-08 12:26:22 -0500 (Fri, 08 May 2009)
```

B.6.3 New files

Purpose. In this section you will make some simple changes: editing an existing file and creating a new file.

From now on all work will happen in the checked out copy of the repository.

Both students should use an editor to create a new file. Make sure to use two distinct names.

```
%% cat > otherfile # both students create a new file
1
2
3
```

Also, let both students edit an existing file, but not the same.

```
%% emacs firstfile # make some changes in an old file, not the same
```

Now use the `svn status` command. Svn will report that it does not know the new file, and that it detects changes in the other.

```
%% svn status
?      otherfile
M      firstfile
```

With `svn add` you tell svn that the new file should become part of the repository:

```
%% svn add otherfile
A         otherfile
%% svn status
A      otherfile
M      firstfile
```

You can now add your changes to the repository:

```
%% svn commit -m "my first batch of changes"
Sending        firstfile
```

```
Adding         otherfile
Transmitting file data ..
Committed revision 3.
```

Since both students have made changes, they need to get those that the other has made:

```
%% svn update
A  mysecondfile
U  myfile
Updated to revision 4.
```

This states that one new file was added, and an existing file updated.

In order for svn to keep track of your files, you should never do `cp` or `mv` on files that are in the repository. Instead, do `svn cp` or `svn mv`. Likewise, there are commands `svn rm` and `svn mkdir`.

B.6.4 Conflicts

Purpose. In this section you will learn about how do deal with conflicting edits by two users of the same repository.

Now let's see what happens if two people edit the same file. Let both students make an edit to `firstfile`, but one to the top, the other to the bottom. After one student commits the edit, the other will see

```
%% emacs firstfile # make some change
%% svn commit -m "another edit to the first file"
Sending        firstfile
svn: Commit failed (details follow):
svn: Out of date: 'firstfile' in transaction '5-1'
```

The solution is to get the other edit, and commit again. After the update, svn reports that it has resolved a conflict successfully.

```
%% svn update
G  firstfile
Updated to revision 5.
%% svn commit -m "another edit to the first file"
Sending        firstfile
Transmitting file data .
Committed revision 6.
```

If both students make edits on the same part of the file, svn can no longer resolve the conflicts. For instance, let one student insert a line between the first and the second, and let the second student edit the second line. Whoever tries to commit second, will get messages like this:

```
%% svn commit -m "another edit to the first file"
svn: Commit failed (details follow):
```

```
svn: Aborting commit: '/share/home/12345/yourname/myproject/firstfile'
remains in conflict
%% svn update
C  firstfile
Updated to revision 7.
```

If you now open the file in an editor, it will look like

```
aa
<<<<<<< .mine
bb
=======
123
b
>>>>>>> .r7
cc
```

indicating the difference between the local version ('mine') and the remote. You need to edit the file to resolve the conflict.

After this, you tell svn that the conflict was resolved, and you can commit:

```
%% svn resolved firstfile
Resolved conflicted state of 'firstfile'
%% svn commit -m "another edit to the first file"
Sending        firstfile
Transmitting file data .
Committed revision 8.
```

The other student then needs to do another update to get the correction.

B.6.5 Inspecting the history

Purpose. In this section, you will learn how to get information about the repository.

You've already seen `svn info` as a way of getting information about the repository. To get the history, do `svn log` to get all log messages, or `svn log 2:5` to get a range.

To see differences in various revisions of individual files, use `svn diff`. First do `svn commit` and `svn update` to make sure you are up to date. Now do `svn diff firstfile`. No output, right? Now make an edit in `firstfile` and do `svn diff firstfile` again. This gives you the difference between the last commited version and the working copy.

You can also ask for differences between committed versions with `svn diff -r 4:6 firstfile`.

The output of this diff command is a bit cryptic, but you can understand it without too much trouble. There are fancy GUI implementations of svn for every platform that show you differences in a much nicer way.

If you simply want to see what a file used to look like, do `svn cat -r 2 firstfile`. To get a copy of a certain revision of the repository, do `svn export -r 3 . ../rev3`.

If you save the output of `svn diff`, it is possible to apply it with the Unix `patch` command. This is a quick way to send patches to someone without them needing to check out the repository.

B.6.6 Shuffling files around

We now realize that we really wanted all these files in a subdirectory in the repository. First we create the directory, putting it under svn control:

```
%% svn mkdir trunk
A         trunk
```

Then we move all files there, again prefixing all comands with `svn`:

```
%% for f in firstfile otherfile myfile mysecondfile ; do \
        svn mv $f trunk/ ; done
A         trunk/firstfile
D         firstfile
A         trunk/otherfile
D         otherfile
A         trunk/myfile
D         myfile
A         trunk/mysecondfile
D         mysecondfile
```

Finally, we commit these changes:

```
%% svn commit -m "trunk created"
Deleting       firstfile
Adding         trunk/firstfile
Deleting       myfile
Deleting       mysecondfile
Deleting       otherfile
Adding         trunk
Adding         trunk/myfile
Adding         trunk/mysecondfile
Adding         trunk/otherfile
```

You probably now have picked up on the message that you always use `svn` to do file manipulations. Let's pretend this has slipped your mind.

- Create a file `somefile` and commit it to the repository.
- Then do `rm firstfile`, thereby deleting a file without `svn` knowing about it.

What is the output of `svn status`?

You can fix this situation in a number of ways:

- `svn revert` restores the file to the state in which it was last restored. For a deleted file, this means that it is brought back into existence from the repository. This command is also useful to undo any local edits, if you change your mind about something.
- `svn rm firstfile` is the official way to delete a file. You can do this even if you have already deleted the file outside `svn`.
- Sometimes `svn` will get confused about your attempts to delete a file. You can then do `svn --force rm yourfile`.

B.6.7 Branching and merging

Suppose you want to tinker with the repository, while still staying up to date with changes that other people make.

```
%% svn copy \
  file:///share/home/12345/you/repository/trunk \
  file:///share/home/12345/you/repository/onebranch \
  -m "create a branch"

Committed revision 11.
```

You can check this out as before:

```
%% svn co file:///share/home/12345/yourname/repository/onebranch \
          projectbranch
A  projectbranch/mysecondfile
A  projectbranch/otherfile
A  projectbranch/myfile
A  projectbranch/firstfile
Checked out revision 11.
```

Now, if you make edits in this branch, they will not be visible in the trunk:

```
%% emacs firstfile # do some edits here
%% svn commit -m "a change in the branch"
Sending        firstfile
Transmitting file data .
Committed revision 13.
%% (cd ../myproject/trunk/ ; svn update )
At revision 13.
```

On the other hand, edits in the main trunk can be pulled into this branch:

```
%% svn merge ^/trunk
--- Merging r13 through r15 into '.':
U    secondfile
```

When you are done editing, the branch edits can be added back to the trunk. For this, it is best to have a clean checkout of the branch:

```
%% svn co  file://‘pwd‘/repository/trunk mycleanproject
A    # all the current files
```

and then do a special merge:

```
%% cd mycleanproject
%% svn merge --reintegrate ^/branch
--- Merging differences between repository URLs into '.':
U    firstfile
 U   .
%% svn info
Path: .
URL: file:///share/home/12345/yourdir/repository/trunk
Repository Root: file:///share/home/12345/yourdir/repository
Repository UUID: dc38b821-b9c6-4a1a-a194-a894fba1d7e7
Revision: 16
Node Kind: directory
Schedule: normal
Last Changed Author: build
Last Changed Rev: 14
Last Changed Date: 2009-05-18 13:34:55 -0500 (Mon, 18 May 2009)
%% svn commit -m "trunk updates from the branch"
Sending        .
Sending        firstfile
Transmitting file data .
Committed revision 17.
```

B.6.8 Repository browsers

The `svn` command can give you some amount of information about a repository, but there are graphical tools that are easier to use and that give a better overview of a repository. For the common platforms, several such tools exist, free or commercial. Here is a browser based tool: `http://wwww.websvn.info`.

B.6.9 Other source control systems

Lately, there has been considerable interest in ‘distributed source code control’ system, such as *git* or *mercurial*. In these systems, each user has a local repository against which commits can be made. This has some advantages over systems like `svn`.

Consider the case where a developer is working on a patch that takes some time to code and test. During this time, development on the project goes on, and by the time the patch can be commited, too many changes have been made for `svn` to be able to merge the patch. As a result, branching and merging is often a frustrating experience with traditional system. In distributed systems, the developer can do commits against the local repository, and roll back

versions as needed. Additionally, when the patch is ready to merge back, it is not a single massive change, but a sequence of small changes, which the source control system should be better able to handle.

A good introduction to Mercurial can be found at `http://www.hginit.com/`.

B.7 Scientific Data Storage

There are many ways of storing data, in particular data that comes in arrays. A surprising number of people stores data in spreadsheets, then exports them to ascii files with comma or tab delimiters, and expects other people (or other programs written by themselves) to read that in again. Such a process is wasteful in several respects:

- The ascii representation of a number takes up much more space than the internal binary representation. Ideally, you would want a file to be as compact as the representation in memory.
- Conversion to and from ascii is slow; it may also lead to loss of precision.

For such reasons, it is desirable to have a file format that is based on binary storage. There are a few more requirements on a useful file format:

- Since binary storage can differ between platforms, a good file format is platform-independent. This will, for instance, prevent the confusion between *big-endian* and *little-endian* storage, as well as conventions of 32 versus 64 bit floating point numbers.
- Application data can be heterogeneous, comprising integer, character, and floating point data. Ideally, all this data should be stored together.
- Application data is also structured. This structure should be reflected in the stored form.
- It is desirable for a file format to be *self-documenting*. If you store a matrix and a right-hand side vector in a file, wouldn't it be nice if the file itself told you which of the stored numbers are the matrix, which the vector, and what the sizes of the objects are?

This tutorial will introduce the HDF5 library, which fulfills these requirements. HDF5 is a large and complicated library, so this tutorial will only touch on the basics. For further information, consult `http://www.hdfgroup.org/HDF5/`. While you do this tutorial, keep your browser open on `http://www.hdfgroup.org/HDF5/RM/RM_H5Front.html` for the exact syntax of the routines.

B.7.1 Introduction to HDF5

As described above, HDF5 is a file format that is machine-independent and self-documenting. Each HDF5 file is set up like a directory tree, with subdirectories, and leaf nodes which contain the actual data. This means that data can be found in a file by referring to its name, rather than its location in the file. In this section you will learn to write programs that write to and read from HDF5 files. In order to check that the files are as you intend, you can use the `h5dump` utility on the command line.[5]

Just a word about compatibility. The HDF5 format is not compatible with the older version HDF4, which is no longer under development. You can still come across people using hdf4 for historic reasons. This tutorial is based on HDF5 version 1.6. Some interfaces changed in the current version 1.8; in order to use 1.6 APIs with 1.8 software, add a flag `-DH5_USE_16_API` to your compile line.

5. In order to do the examples, the `h5dump` utility needs to be in your path, and you need to know the location of the `hdf5.h` and `libhdf5.a` and related library files.

B.7.2 Creating a file

First of all, we need to create an HDF5 file.

```
file_id = H5Fcreate( filename, ... );
...
status = H5Fclose(file_id);
```

This file will be the container for a number of data items, organized like a directory tree.

Exercise. Create an HDF5 file by compiling and running the `create.c` example below.

Expected outcome. A file `file.h5` should be created.

Caveats. Be sure to add HDF5 include and library directories:
`cc -c create.c -I. -I/opt/local/include`
and
`cc -o create create.o -L/opt/local/lib -lhdf5`. The include and lib directories will be system dependent.

```
#include "myh5defs.h"
#define FILE "file.h5"

main() {

   hid_t       file_id;   /* file identifier */
   herr_t      status;

   /* Create a new file using default properties. */
   file_id = H5Fcreate(FILE, H5F_ACC_TRUNC, H5P_DEFAULT, H5P_DEFAULT); H5REPORT(fi

   /* Terminate access to the file. */
   status = H5Fclose(file_id);
}
```

You can display the created file on the commandline:

```
%% h5dump file.h5
HDF5 "file.h5" {
GROUP "/" {
}
}
```

Note that an empty file corresponds to just the root of the directory tree that will hold the data.

B.7.3 Datasets

Next we create a dataset, in this example a 2D grid. To describe this, we first need to construct a dataspace:

```
dims[0] = 4; dims[1] = 6;
dataspace_id = H5Screate_simple(2, dims, NULL);
dataset_id = H5Dcreate(file_id, "/dset", dataspace_id, .... );
....
status = H5Dclose(dataset_id);
status = H5Sclose(dataspace_id);
```

Note that datasets and dataspaces need to be closed, just like files.

Exercise. Create a dataset by compiling and running the `dataset.c` code below

Expected outcome. This creates a file `dset.h` that can be displayed with `h5dump`.

```
#include "myh5defs.h"
#define FILE "dset.h5"

main() {

   hid_t       file_id, dataset_id, dataspace_id;  /* identifiers */
   hsize_t     dims[2];
   herr_t      status;

   /* Create a new file using default properties. */
   file_id = H5Fcreate(FILE, H5F_ACC_TRUNC, H5P_DEFAULT, H5P_DEFAULT);

   /* Create the data space for the dataset. */
   dims[0] = 4;
   dims[1] = 6;
   dataspace_id = H5Screate_simple(2, dims, NULL);

   /* Create the dataset. */
   dataset_id = H5Dcreate(file_id, "/dset", H5T_NATIVE_INT, dataspace_id, H5P_DEFAULT);
   /*H5T_STD_I32BE*/

   /* End access to the dataset and release resources used by it. */
   status = H5Dclose(dataset_id);

   /* Terminate access to the data space. */
   status = H5Sclose(dataspace_id);
```

```
   /* Close the file. */
   status = H5Fclose(file_id);
}
```

We again view the created file online:

```
%% h5dump dset.h5
HDF5 "dset.h5" {
GROUP "/" {
   DATASET "dset" {
      DATATYPE  H5T_STD_I32BE
      DATASPACE  SIMPLE { ( 4, 6 ) / ( 4, 6 ) }
      DATA {
      (0,0): 0, 0, 0, 0, 0, 0,
      (1,0): 0, 0, 0, 0, 0, 0,
      (2,0): 0, 0, 0, 0, 0, 0,
      (3,0): 0, 0, 0, 0, 0, 0
      }
   }
}
}
```

The datafile contains such information as the size of the arrays you store. Still, you may want to add related scalar information. For instance, if the array is output of a program, you could record with what input parameter was it generated.

```
   parmspace = H5Screate(H5S_SCALAR);
   parm_id = H5Dcreate
     (file_id,"/parm",H5T_NATIVE_INT,parmspace,H5P_DEFAULT);
```

Exercise. Add a scalar dataspace to the HDF5 file, by compiling and running the `parmwrite.c` code below.

Expected outcome. A new file `wdset.h5` is created.

```
#include "myh5defs.h"
#define FILE "pdset.h5"

main() {

   hid_t       file_id, dataset_id, dataspace_id;  /* identifiers */
   hid_t       parm_id,parmspace;
   hsize_t     dims[2];
   herr_t      status;
```

```
   /* Create a new file using default properties. */
   file_id = H5Fcreate(FILE, H5F_ACC_TRUNC, H5P_DEFAULT, H5P_DEFAULT);

   /* Create the data space for the dataset. */
   dims[0] = 4;
   dims[1] = 6;
   dataspace_id = H5Screate_simple(2, dims, NULL);

   /* Create the dataset. */
   dataset_id = H5Dcreate
     (file_id, "/dset", H5T_STD_I32BE, dataspace_id, H5P_DEFAULT);

   /* Add a descriptive parameter */
   parmspace = H5Screate(H5S_SCALAR);
   parm_id = H5Dcreate
     (file_id,"/parm",H5T_NATIVE_INT,parmspace,H5P_DEFAULT);

   /* End access to the dataset and release resources used by it. */
   status = H5Dclose(dataset_id);
   status = H5Dclose(parm_id);

   /* Terminate access to the data space. */
   status = H5Sclose(dataspace_id);
   status = H5Sclose(parmspace);

   /* Close the file. */
   status = H5Fclose(file_id);
}


%% h5dump wdset.h5
HDF5 "wdset.h5" {
GROUP "/" {
   DATASET "dset" {
      DATATYPE  H5T_IEEE_F64LE
      DATASPACE  SIMPLE { ( 4, 6 ) / ( 4, 6 ) }
      DATA {
      (0,0): 0.5, 1.5, 2.5, 3.5, 4.5, 5.5,
      (1,0): 6.5, 7.5, 8.5, 9.5, 10.5, 11.5,
      (2,0): 12.5, 13.5, 14.5, 15.5, 16.5, 17.5,
      (3,0): 18.5, 19.5, 20.5, 21.5, 22.5, 23.5
```

```
      }
   }
   DATASET "parm" {
      DATATYPE  H5T_STD_I32LE
      DATASPACE  SCALAR
      DATA {
      (0): 37
      }
   }
}
}
```

B.7.4 Writing the data

The datasets you created allocate the space in the hdf5 file. Now you need to put actual data in it. This is done with the `H5Dwrite` call.

```
/* Write floating point data */
for (i=0; i<24; i++) data[i] = i+.5;
status = H5Dwrite
  (dataset,H5T_NATIVE_DOUBLE,H5S_ALL,H5S_ALL,H5P_DEFAULT,
   data);
/* write parameter value */
parm = 37;
status = H5Dwrite
  (parmset,H5T_NATIVE_INT,H5S_ALL,H5S_ALL,H5P_DEFAULT,
   &parm);
```

```
#include "myh5defs.h"
#define FILE "wdset.h5"

main() {

   hid_t       file_id, dataset, dataspace;  /* identifiers */
   hid_t       parmset,parmspace;
   hsize_t     dims[2];
   herr_t      status;
   double data[24]; int i,parm;

   /* Create a new file using default properties. */
   file_id = H5Fcreate(FILE, H5F_ACC_TRUNC, H5P_DEFAULT, H5P_DEFAULT);
```

```
   /* Create the dataset. */
   dims[0] = 4; dims[1] = 6;
   dataspace = H5Screate_simple(2, dims, NULL);
   dataset = H5Dcreate
     (file_id, "/dset", H5T_NATIVE_DOUBLE, dataspace, H5P_DEFAULT);

   /* Add a descriptive parameter */
   parmspace = H5Screate(H5S_SCALAR);
   parmset = H5Dcreate
     (file_id,"/parm",H5T_NATIVE_INT,parmspace,H5P_DEFAULT);

   /* Write data to file */
   for (i=0; i<24; i++) data[i] = i+.5;
   status = H5Dwrite
     (dataset,H5T_NATIVE_DOUBLE,H5S_ALL,H5S_ALL,H5P_DEFAULT,
      data); H5REPORT(status);

   /* write parameter value */
   parm = 37;
   status = H5Dwrite
     (parmset,H5T_NATIVE_INT,H5S_ALL,H5S_ALL,H5P_DEFAULT,
      &parm); H5REPORT(status);

   /* End access to the dataset and release resources used by it. */
   status = H5Dclose(dataset);
   status = H5Dclose(parmset);

   /* Terminate access to the data space. */
   status = H5Sclose(dataspace);
   status = H5Sclose(parmspace);

   /* Close the file. */
   status = H5Fclose(file_id);
}


%% h5dump wdset.h5
HDF5 "wdset.h5" {
GROUP "/" {
   DATASET "dset" {
      DATATYPE  H5T_IEEE_F64LE
```

```
      DATASPACE  SIMPLE { ( 4, 6 ) / ( 4, 6 ) }
      DATA {
      (0,0): 0.5, 1.5, 2.5, 3.5, 4.5, 5.5,
      (1,0): 6.5, 7.5, 8.5, 9.5, 10.5, 11.5,
      (2,0): 12.5, 13.5, 14.5, 15.5, 16.5, 17.5,
      (3,0): 18.5, 19.5, 20.5, 21.5, 22.5, 23.5
      }
   }
   DATASET "parm" {
      DATATYPE  H5T_STD_I32LE
      DATASPACE  SCALAR
      DATA {
      (0): 37
      }
   }
}
}
```

If you look closely at the source and the dump, you see that the data types are declared as 'native', but rendered as `LE`. The 'native' declaration makes the datatypes behave like the built-in C or Fortran data types. Alternatively, you can explicitly indicate whether data is *little endian* or *big endian*. These terms describe how the bytes of a data item are ordered in memory. Most architectures use little endian, as you can see in the dump output, but, notably, IBM uses big endian.

B.7.5 Reading

Now that we have a file with some data, we can do the mirror part of the story: reading from that file. The essential commands are

```
h5file = H5Fopen( .... )
....
H5Dread( dataset, .... data .... )
```

where the `H5Dread` command has the same arguments as the corresponding `H5Dwrite`.

Exercise. Read data from the `wdset.h5` file that you create in the previous exercise, by compiling and running the `allread.c` example below.

Expected outcome. Running the `allread` executable will print the value `37` of the parameter, and the value `8.5` of the `(1,2)` data point of the array.

Caveats. Make sure that you run `parmwrite` to create the input file.

```
 %% ./allread
parameter value: 37
arbitrary data point [1,2]: 8.500000e+00
```

B.8 Scientific Libraries

There are many libraries for scientific computing. Some are specialized to certain application areas, others are quite general. In this section we will take a brief look at the PETSc library for sparse matrix computations, and the BLAS/Lapack libraries for dense computations.

B.8.1 The Portable Extendable Toolkit for Scientific Computing

PETSc, the Portable Extendable Toolkit for Scientifc Computation [], is a large powerful library, mostly concerned with linear and nonlinear system of equations that arise from discretized PDEs. Since it has many hundreds of routines (and a good manual already exists) we limit this tutorial to a run-through of a simple, yet representative, PETSc program.

B.8.1.1 What is in PETSc?

PETSc can be used as a library in the traditional sense, where you use some high level functionality, such as solving a nonlinear system of equations, in your program. However, it can also be used as a toolbox, to compose your own numerical applications using low-level tools.

- Linear system solvers (sparse/dense, iterative/direct)
- Nonlinear system solvers
- Tools for distributed matrices
- Support for profiling, debugging, graphical output

The basic functionality of PETSc can be extended through external packages:

- Dense linear algebra: `Scalapack`, `Plapack`
- Grid partitioning software: `ParMetis`, `Jostle`, `Chaco`, `Party`
- ODE solvers: `PVODE`
- Eigenvalue solvers (including SVD): `SLEPc`
- Optimization: `TAO`

B.8.1.2 Design of PETSc

All objects in Petsc are defined on an MPI communicator (section 2.5.2. This way, you can define a matrix or a linear system on all your available processors, or on a subset. PETSc objects can only interact if they are based on the same communicator.

Parallelism is handled through MPI. You may need to have occasional MPI calls in a PETSc code, but the vast bulk of communication is done behind the scenes inside PETSc calls. Shared memory parallel (such as through OpenMP; section 2.5.1) is not used explicitly, but the user can incorporate it without problems.

PETSc is object-oriented in design, although it is coded in standard C. For instance, a call

```
MATMult(A,x,y); // y <- A x
```

looks the same, regardless whether A is sparse or dense, sequential or parallel.

One side effect of this is that the actual data are usually hidden from the user. Data can be accessed through routines such as

```
double *array;
VecGetArray(myvector,&array);
```

but most of the time the application does not explicitly maintain the data, only the PETSc objects around them. As the ultimate consequence of this, the user does not allocate data; rather, matrix and vector arrays get created through the PETSc create calls, and subsequent values are inserted through PETSc calls:

```
MatSetValue(A,i,j,v,INSERT_VALUES); // A[i,j] <- v
```

This may feel like the user is giving up control, but it actually makes programming a lot easier, since the user does not have to worry about details of storage schemes, especially in parallel.

B.8.1.3 An example program

We will now go through a PETSc example program that solves a linear system by an iterative method. This example is intended to convey the 'taste' of PETSc, not all the details. (Note: the PETSc source tree comes with many examples.)

B.8.1.3.1 **Construction of the coefficient matrix** We will construct the matrix of 5-point stencil for the Poisson operator (section 4.2.2.2) as a sparse matrix.

We start by determining the size of the matrix. PETSc has an elegant mechanism for getting parameters from the program commandline. This prevents the need for recompilation of the code.

C:

```
int domain_size;
PetscOptionsGetInt
  (PETSC_NULL,"-n",&domain_size,&flag);
if (!flag) domain_size = 10;
matrix_size = domain_size*domain_size;
```

Fortran:

```
call PetscOptionsGetInt(PETSC_NULL_CHARACTER,
>     "-n",domain_size,flag)
if (.not.flag) domain_size = 10
matrix_size = domain_size*domain_size;
```

Creating a matrix involves specifying its communicator, its type (here: distributed sparse), and its local and global sizes. In this case, we specify the global size and leave the local sizes and data distribution up to PETSc.

C:

```
comm = MPI_COMM_WORLD; // or subset
MatCreate(comm,&A);
MatSetType(A,MATMPIAIJ);
MatSetSizes(A,PETSC_DECIDE,PETSC_DECIDE,
  matrix_size,matrix_size);
```

Fortran:

```
comm = MPI_COMM_WORLD ! or subset
call MatCreate(comm,A)
call MatSetType(A,MATMPIAIJ)
call MatSetSizes(A,PETSC_DECIDE,PETSC_DECIDE,
>     matrix_size,matrix_size)
```

Both matrices and vectors (see below) have a local and a global size. The local part is a contiguous block of vector entries or matrix rows. For matrix-vector operations, the matrix and vector need to be have compatible local sizes. It is often easiest to specify only the global size and let PETSc determine the local size; on the other hand, if the number of local variables is determined by the application, PETSc will add the local sizes to determine the global size.

B.8.1.3.2 **Filling in matrix elements** We will now set matrix elements (refer to figure 4.1) by having each processor iterate over the full domain, but only inserting those elements that are in its matrix block row.

C:

```
MatGetOwnershipRange(A,&low,&high);
for ( i=0; i<m; i++ ) {
  for ( j=0; j<n; j++ ) {
    I = j + n*i;
    if (I>=low && I<high) {
      J = I-1; v = -1.0;
      if (i>0) MatSetValues
          (A,1,&I,1,&J,&v,INSERT_VALUES);
      J = I+1 // et cetera
    }
  }
}
MatAssemblyBegin(A,MAT_FINAL_ASSEMBLY);
MatAssemblyEnd(A,MAT_FINAL_ASSEMBLY);
```

Fortran:

```
call MatGetOwnershipRange(A,low,high)
do i=0,m-1
   do j=0,n-1
```

```
        ii = j + n*i
        if (ii>=low .and. ii<high) then
           jj = ii - n
           if (j>0) call MatSetValues
>                     (A,1,ii,1,jj,v,INSERT_VALUES)
           ...
  call MatAssemblyBegin(A,MAT_FINAL_ASSEMBLY)
  call MatAssemblyEnd(A,MAT_FINAL_ASSEMBLY)
```

B.8.1.3.3 **Finite Element Matrix assembly** PETSc's versatility in dealing with Finite Element matrices (see sections 4.2.2.4 and 6.6.2), where elements are constructed by adding together contributions, sometimes from different processors. This is no problem in PETSc: any processor can set (or add to) any matrix element. The assembly calls will move data to their eventual location on the correct processors.

```
for (e=myfirstelement; e<mylastelement; e++) {
  for (i=0; i<nlocalnodes; i++) {
    I = localtoglobal(e,i);
    for (j=0; j<nlocalnodes; j++) {
      J = localtoglobal(e,j);
      v = integration(e,i,j);
      MatSetValues
          (mat,1,&I,1,&J,&v,ADD_VALUES);
      ....
    }
  }
}
MatAssemblyBegin(mat,MAT_FINAL_ASSEMBLY);
MatAssemblyEnd(mat,MAT_FINAL_ASSEMBLY);
```

B.8.1.3.4 **Linear system solver** Next we declare an iterative method and preconditioner to solve the linear system with. PETSc has a large repertoire of Krylov methods (section 5.5.7) and preconditioners.

C:

```
KSPCreate(comm,&Solver);
KSPSetOperators(Solver,A,A,0);
KSPSetType(Solver,KSPCGS);
{
  PC Prec;
  KSPGetPC(Solver,&Prec);
  PCSetType(Prec,PCJACOBI);
}
```

Fortran:

```
call KSPCreate(comm,Solve)
call KSPSetOperators(Solve,A,A,0)
call KSPSetType(Solve,KSPCGS)
call KSPGetPC(Solve,Prec)
call PCSetType(Prec,PCJACOBI)
```

The actual preconditioner is not yet created: that will happen in the solve call.

B.8.1.3.5 **Input and output vectors** Finally, we need our right hand side and solution vectors. To make them compatible with the matrix, they need to be created with the same local and global sizes. This can be done by literally specifying the same parameters, but a more elegant way is to query the matrix size and use the result.

C:

```
MatGetLocalSize(A,&isize,PETSC_NULL); // get local size
VecCreateMPI(comm,isize,PETSC_DECIDE,&Rhs);
// explicit alternative:
// VecCreateMPI(comm,PETSC_DECIDE,matrix_size,&Rhs);
VecDuplicate(Rhs,&Sol);
VecSet(Rhs,one);
```

Fortran:

```
call MatGetLocalSize(A,isize,PETSC_NULL_INTEGER)
call VecCreateMPI(comm,isize,PETSC_DECIDE,Rhs)
call VecDuplicate(Rhs,Sol)
call VecSet(Rhs,one)
```

B.8.1.3.6 **Solving the system** Solving the linear system is a one line call to `KSPSolve`. The story would end there if it weren't for some complications:

- Iterative methods can fail, and the solve call does not tell us whether that happened.
- If the system was solved successfully, we would like to know in how many iterations.
- There can be other reason for the iterative method to halt, such as reaching its maximum number of iterations without converging.

The following code snipped illustrates how this knowledge can be extracted from the solver object.

C:

```
KSPSolve(Solver,Rhs,Sol)
{
  PetscInt its; KSPConvergedReason reason;
  Vec Res; PetscReal norm;
  KSPGetConvergedReason(Solver,&reason);
```

```
    if (reason<0) {
      PetscPrintf(comm,"Failure to converge\n");
    } else {
      KSPGetIterationNumber(Solver,&its);
      PetscPrintf(comm,"Number of iterations: %d\n",its);
    }
  }
```

Fortran:

```
  call KSPSolve(Solve,Rhs,Sol)

  call KSPGetConvergedReason(Solve,reason)
  if (reason<0) then
     call PetscPrintf(comm,"Failure to converge\n")
  else
     call KSPGetIterationNumber(Solve,its)
     write(msg,10) its
10   format('Number of iterations: i4")
     call PetscPrintf(comm,msg)
  end if
```

B.8.1.3.7 **Further operations** As an illustration of the toolbox nature of PETSc, here is a code snippet that computes the residual vector after a linear system solution. This operation could have been added to the library, but composing it yourself offers more flexibility.

C:

```
  VecDuplicate(Rhs,&Res);
  MatMult(A,Sol,Res);
  VecAXPY(Res,-1,Rhs);
  VecNorm(Res,NORM_2,&norm);
  PetscPrintf(MPI_COMM_WORLD,"residual norm: %e\n",norm);
```

Fortran:

```
  call VecDuplicate(Rhs,Res)
  call MatMult(A,Sol,Res)
  call VecAXPY(Res,mone,Rhs)
  call VecNorm(Res,NORM_2,norm)
  if (mytid==0) print *,"residual norm:",norm
```

Finally, we need to free all objects, both the obvious ones, such as the matrix and the vectors, and the non-obvious ones, such as the solver.

```
  VecDestroy(Res);
```

```
KSPDestroy(Solver);
VecDestroy(Rhs);
VecDestroy(Sol);
```

```
call MatDestroy(A)
call KSPDestroy(Solve)
call VecDestroy(Rhs)
call VecDestroy(Sol)
call VecDestroy(Res)
```

B.8.1.4 Quick experimentation

Reading a parameter from the commandline above is actually a special case of a general mechanism for influencing PETSc's workings through commandline options.

Here is an example of setting the iterative solver and preconditioner from the commandline:

```
yourprog -ksp_type gmres -ksp_gmres_restart 25
    -pc_type ilu -pc_factor_levels 3
```

In order for this to work, your code needs to call

```
KSPSetFromOptions(solver);
```

before the system solution. This mechanism is very powerful, and it obviates the need for much code recompilation.

B.8.2 Libraries for dense linear algebra: Lapack and Scalapack

Dense linear algebra, that is linear algebra on matrices that are stored as two-dimensional arrays (as opposed to sparse linear algebra; see sections 5.4.1 and 5.4, as well as the tutorial on PETSc B.8) has been standardized for a considerable time. The basic operations are defined by the three levels of *Basic Linear Algebra Subprograms (BLAS)*:

- Level 1 defines vector operations that are characterized by a single loop [61].
- Level 2 defines matrix vector operations, both explicit such as the matrix-vector product, and implicit such as the solution of triangular systems [28].
- Level 3 defines matrix-matrix operations, most notably the matrix-matrix product [27].

Based on these building blocks libraries have been built that tackle the more sophisticated problems such as solving linear systems, or computing eigenvalues or singular values. *Linpack*[6] and Eispack were the first to formalize these operations involved, using Blas Level 1 and Blas Level 2 respectively. A later development, *Lapack* uses the blocked operations of Blas Level 3. As you saw in section 1.4.1, this is needed to get high performance on cache-based CPUs. (Note: the reference implementation of the BLAS [15] will not give good performance with any compiler; most platforms have vendor-optimized implementations, such as the *MKL* library from Intel.)

6. The linear system solver from this package later became the *Linpack benchmark*.

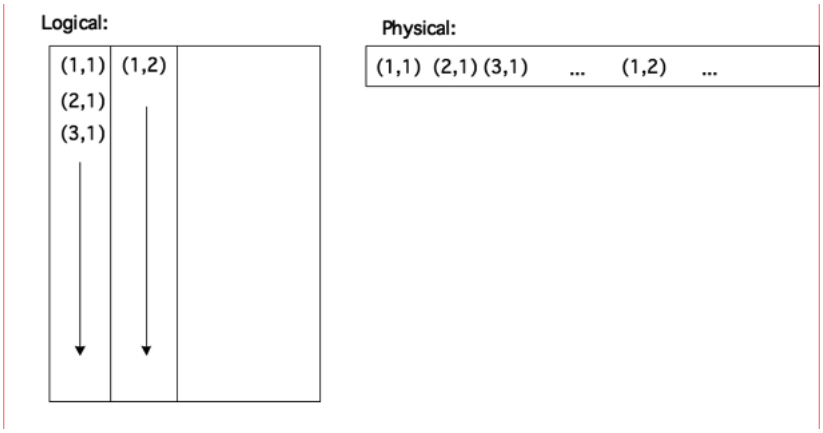


Figure B.2: Column-major storage of an array in Fortran

With the advent of parallel computers, several projects arose that extended the Lapack functionality to distributed computing, most notably *Scalapack* [20] and *PLapack* [79, 78]. These packages are considerably harder to use than Lapack[7] because of the need for the two-dimensional block cyclic distribution; sections 6.3 and 6.4. We will not go into the details here.

B.8.2.1 BLAS matrix storage

There are a few points to bear in mind about the way matrices are stored in the BLAS and LAPACK[8]:

- Since these libraries originated in a Fortran environment, they use 1-based indexing. Users of languages such as C/C++ are only affected by this when routines use index arrays, such as the location of pivots in LU factorizations.
- Columnwise storage
- Leading dimension

B.8.2.1.1 **Fortran column-major ordering** Since computer memory is one-dimensional, some conversion is needed from two-dimensional matrix coordinates to memory locations. The Fortran language uses *column-major storage*, that is, elements in a column are stored consecutively; see figure B.2. This is also described informally as 'the leftmost index varies quickest'.

B.8.2.1.2 **Submatrices and the `LDA` parameter** Using the storage scheme described above, it is clear how to store an $m \times n$ matrix in mn memory locations. However, there are many cases where software needs access to a matrix that is a subblock of another, larger, matrix. As you see in figure B.3 such a subblock is no longer

7. PLapack is probably the easier to use of the two.
8. We are not going into band storage here.

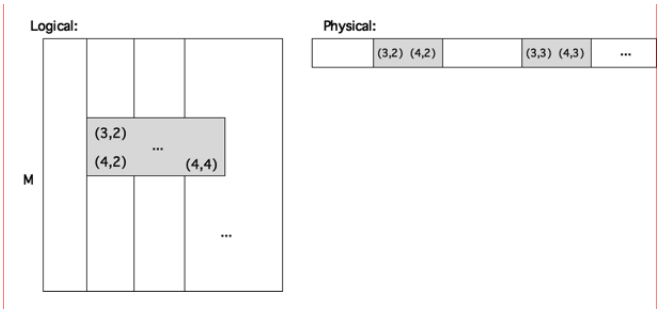


Figure B.3: A subblock out of a larger matrix

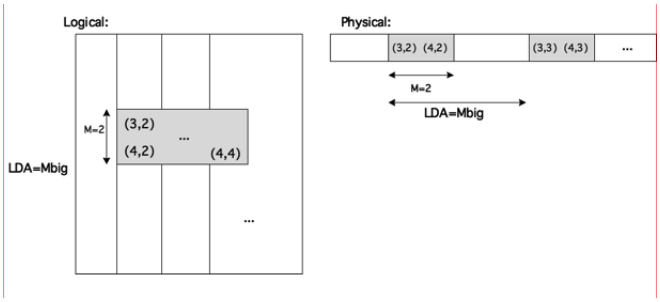


Figure B.4: A subblock out of a larger matrix

contiguous in memory. The way to describe this is by introducing a third parameter in addition to `M,N`: we let `LDA` be the 'leading dimension of `A`', that is, the allocated first dimension of the surrounding array. This is illustrated in figure B.4.

B.8.2.2 Organisation of routines

Lapack is organized with three levels of routines:

- Drivers. These are powerful top level routine for problems such as solving linear systems or computing an SVD. There are simple and expert drivers; the expert ones have more numerical sophistication.
- Computational routines. These are the routines that drivers are built up out of[9]. A user may have occasion to call them by themselves.
- Auxiliary routines.

Routines conform to a general naming scheme: `XYYZZZ` where

X precision: `S,D,C,Z` stand for single and double, single complex and double complex, respectively.
YY storage scheme: general rectangular, triangular, banded.
ZZZ operation. See the manual for a list.

9. Ha! Take that, Winston.

Expert driver names end on 'X'.

B.8.2.2.1 Lapack data formats

28 formats, including

GE General matrix: store `A(LDA,*)`

SY/HE Symmetric/Hermitian: general storage; `UPLO` parameter to indicate upper or lower (e.g. `SPOTRF`)

GB/SB/HB General/symmetric/Hermitian band; these formats use column-major storage; in `SGBTRF` overallocation needed because of pivoting

PB Symmetric of Hermitian positive definite band; no overallocation in `SPDTRF`

B.8.2.2.2 Lapack operations

- Linear system solving. Simple drivers: `-SV` (e.g., `DGESV`) Solve $AX = B$, overwrite A with LU (with pivoting), overwrite B with X.
  Expert driver: `-SVX` Also transpose solve, condition estimation, refinement, equilibration
- Least squares problems. Drivers:
  `xGELS` using QR or LQ under full-rank assumption
  `xGELSY` "complete orthogonal factorisation"
  `xGELSS` using SVD
  `xGELSD` using divide-conquer SVD (faster, but more workspace than `xGELSS`)
  Also: LSE & GLM linear equality constraint & general linear model
- Eigenvalue routines. Symmetric/Hermitian: `xSY` or `xHE` (also `SP`, `SB`, `ST`) simple driver `-EV` expert driver `-EVX` divide and conquer `-EVD` relative robust representation `-EVR`
  General (only `xGE`) Schur decomposition `-ES` and `-ESX` eigenvalues `-EV` and `-EVX`
  SVD (only `xGE`) simple driver `-SVD` divide and conquer `SDD`
  Generalized symmetric (`SY` and `HE`; `SP`, `SB`) simple driver `GV` expert `GVX` divide-conquer `GVD`
  Nonsymmetric Schur: simple `GGES`, expert `GGESX` eigen: simple `GGEV`, expert `GGEVX`
  svd: `GGSVD`

B.9 Plotting with GNUplot

The `gnuplot` utility is a simple program for plotting sets of points or curves. This very short tutorial will show you some of the basics. For more commands and options, see the manual `http://www.gnuplot.info/docs/gnuplot.html`.

B.9.1 Usage modes

The two modes for running `gnuplot` are *interactive* and *from file*. In interactive mode, you call `gnuplot` from the command line, type commands, and watch output appear (see next paragraph); you terminate an interactive session with `quit`. If you want to save the results of an interactive session, do `save "name.plt"`. This file can be edited, and loaded with `load "name.plt"`.

Plotting non-interactively, you call `gnuplot <your file>`.

The output of `gnuplot` can be a picture on your screen, or drawing instructions in a file. Where the output goes depends on the setting of the *terminal*. By default, `gnuplot` will try to draw a picture. This is equivalent to declaring

```
set terminal x11
```

or `aqua`, `windows`, or any choice of graphics hardware.

For output to file, declare

```
set terminal pdf
```

or `fig`, `latex`, `pbm`, et cetera. Note that this will only cause the pdf commands to be written to your screen: you need to direct them to file with

```
  set ouput "myplot.pdf"
```

or capture them with

```
  gnuplot my.plt > myplot.pdf
```

B.9.2 Plotting

The basic plot commands are `plot` for 2D, and `splot` ('surface plot') for 3D plotting.

B.9.2.1 Plotting curves

By specifying

```
plot x**2
```

you get a plot of $f(x) = x^2$; `gnuplot` will decide on the range for x. With

```
set xrange [0:1]
plot 1-x title "down", x**2 title "up"
```

you get two graphs in one plot, with the x range limited to $[0, 1]$, and the appropriate legends for the graphs. The variable `x` is the default for plotting functions.

Plotting one function against another – or equivalently, plotting a parametric curve – goes like this:

```
set parametric
plot [t=0:1.57] cos(t),sin(t)
```

which gives a quarter circle.

To get more than one graph in a plot, use the command `set multiplot`.

B.9.2.2 Plotting data points

It is also possible to plot curves based on data points. The basic syntax is `plot 'datafile'`, which takes two columns from the data file and interprets them as (x, y) coordinates. Since data files can often have multiple columns of data, the common syntax is `plot 'datafile' using 3:6` for columns 3 and 6. Further qualifiers like `with lines` indicate how points are to be connected.

Similarly, `splot "datafile3d.dat" 2:5:7` will interpret three columns as specifying (x, y, z) coordinates for a 3D plot.

If a data file is to be interpreted as level or height values on a rectangular grid, do `splot "matrix.dat" matrix` for data points; connect them with

```
  split "matrix.dat" matrix with lines
```

B.9.2.3 Customization

Plots can be customized in many ways. Some of these customizations use the `set` command. For instance,

```
  set xlabel "time"
  set ylabel "output"
  set title "Power curve"
```

You can also change the default drawing style with

```
set style function dots
```

(`dots`, `lines`, `dots`, `points`, et cetera), or change on a single plot with

```
plot f(x) with points
```

B.9.3 Workflow

Imagine that your code produces a dataset that you want to plot, and you run your code for a number of inputs. It would be nice if the plotting can be automated. Gnuplot itself does not have the facilities for this, but with a little help from shell programming this is not hard to do.

Suppose you have data files

```
  data1.dat data2.dat data3.dat
```

and you want to plot them with the same gnuplot commands. You could make a file `plot.template`:

```
set term pdf
set output "FILENAME.pdf"
plot "FILENAME.dat"
```

The string `FILENAME` can be replaced by the actual file names using, for instance `sed`:

```
for d in data1 data2 data3 ; do
  cat plot.template | sed s/FILENAME/$d/ > plot.cmd
  gnuplot plot.cmd
done
```

Variations on this basic idea are many.

B.10 Programming languages

B.10.1 C/Fortran interoperability

Most of the time, a program is written is written in a single language, but in some circumstances it is necessary or desirable to mix sources in more than one language for a single executable. One such case is when a library is written in one language, but used by a program in another. In such a case, the library writer will probably have made it easy for you to use the library; this section is for the case that you find yourself in this situation.

B.10.1.1 Arrays

C and Fortran have different conventions for storing multi-dimensional arrays. You need to be aware of this when you pass an array between routines written in different languages.

Fortran stores multi-dimensional arrays in *column-major* order. For two dimensional arrays (`A(i,j)`) this means that the elements in each column are stored contiguously: a 2×2 array is stored as `A(1,1), A(2,1), A(1,2), A(2,2)`. Three and higher dimensional arrays are an obvious extension: it is sometimes said that 'the left index varies quickest'.

C arrays are stored in *row-major* order: elements in each row are stored contiguous, and columns are then placed sequentially in memory. A 2×2 array `A[2][2]` is then stored as `A[1][1], A[1][2], A[2][1], A[2][2]`.

A number of remarks about arrays in C.

- C (before the C99 standard) has multi-dimensional arrays only in a limited sense. You can declare them, but if you pass them to another C function, they no longer look multi-dimensional: they have become plain `float*` (or whatever type) arrays. That brings me to the next point.
- Multi-dimensional arrays in C look as if they have type `float**`, that is, an array of pointers that point to (separately allocated) arrays for the rows. While you could certainly implement this:
  ```
  float **A;
  A = (float**)malloc(m*sizeof(float*));
  for (i=0; i<n; i++)
    A[i] = (float*)malloc(n*sizeof(float));
  ```
  careful reading of the standard reveals that a multi-dimensional array is in fact a single block of memory, no further pointers involved.

Given the above limitation on passing multi-dimensional arrays, and the fact that a C routine can not tell whether it's called from Fortran or C, it is best not to bother with multi-dimensional arrays in C, and to emulate them:

```
float *A;
A = (float*)malloc(m*n*sizeof(float));
#define sub(i,j,m,n) i+j*m
for (i=0; i<m; i++)
  for (j=0; j<n; j++)
    .... A[sub(i,j,m,n)] ....
```

where for interoperability we store the elements in column-major fashion.

B.10.1.2 Strings

Programming languages differ widely in how they handle strings.

- In C, a string is an array of characters; the end of the string is indicated by a null character, that is the ascii character zero, which has an all zero bit pattern. This is called *null termination*.
- In Fortran, a string is an array of characters. The length is maintained in a internal variable, which is passed as a hidden parameter to subroutines.
- In Pascal, a string is an array with an integer denoting the length in the first position. Since only one byte is used for this, strings can not be longer than 255 characters in Pascal.

As you can see, passing strings between different languages is fraught with peril. The safest solution is to use null-terminated strings throughout; some compilers support extensions to facilitate this, for instance writing

```
DATA forstring /'This is a null-terminated string.'C/
```

Recently, the 'C/Fortran interoperability standard' has provided a systematic solution to this.

B.10.1.3 Subprogram arguments

In C, you pass a `float` argument to a function if the function needs its value, and `float*` if the function has to modify the value of the variable in the calling environment. Fortran has no such distinction: every variable is passed *by reference*. This has some strange consequences: if you pass a literal value (`37.5`) to a subroutine, the compiler will allocate a nameless variable with that value, and pass the address of it, rather than the value[10].

For interfacing Fortran and C routines, this means that a Fortran routine looks to a C program like all its argument are 'star' arguments. Conversely, if you want a C subprogram to be callable from Fortran, all its arguments have to be star-this or that. This means on the one hand that you will sometimes pass a variable by reference that you would like to pass by value.

Worse, it means that C subprograms like

```
void mysub(int **i) {
 *i = (int*)malloc(8*sizeof(int));
 return;
}
```

can not be called from Fortran. There is a hack to get around this (check out the Fortran77 interface to the Petsc routine `VecGetValues`) and with more cleverness you can use `POINTER` variables for this.

B.10.1.4 Linker conventions

As explained above, a compiler turns a source file into a binary, which no longer has any trace of the source language: it contains in effect functions in assembly language. The linker will then match up calls and definitions. The problem with using multiple languages is then that compilers have different notions of how to translate function names from the source file to the binary file.

---

10. With a bit of cleverness and the right compiler, you can have a program that says `print *,7` and prints `8` because of this.

The most common case of language interoperability is between C and Fortran. The problems are platform dependent, but commonly

- The Fortran compiler attaches a trailing underscore to function names in the object file.
- The C compiler takes the function name as it is in the source.

Since C is a popular language to write libraries in, this means we can solve the problem by either

- Appending an underscore to all C function names; or
- Include a simple wrapper call:

```
int SomeCFunction(int i,float f)
{
  .....
}
int SomeCFunction_(int i,float f)
{
  return SomeCFunction(i,f);
}
```

With the latest Fortran standard it is possible to declare the external name of variables and routines:

```
%% cat t.f
module operator
  real, bind(C) :: x
contains
  subroutine s(), bind(C,name='_s')
  return
  end subroutine
  ...
end module
%% ifort -c t.f
%% nm t.o
.... T _s
.... C _x
```

It is also possible to declare data types to be C-compatible:

```
  type, bind(C) :: c_comp
    real (c_float)  :: data
    integer (c_int) :: i
    type (c_ptr)    :: ptr
  end type
```

B.10.1.5 Input/output

Both languages have their own system for handling input/output, and it is not really possible to meet in the middle. Basically, if Fortran routines do I/O, the main program has to be in Fortran. Consequently, it is best to isolate I/O as

much as possible, and use C for I/O in mixed language programming.

B.10.1.6 Fortran/C interoperability in Fortran2003

The latest version of Fortran, unsupported by many compilers at this time, has mechanisms for interfacing to C.

- There is a module that contains named kinds, so that one can declare
  `INTEGER,KIND(C_SHORT) :: i`
- Fortran pointers are more complicated objects, so passing them to C is hard; Fortran2003 has a mechanism to deal with C pointers, which are just addresses.
- Fortran derived types can be made compatible with C structures.

Appendix C

Class project

C.1 Heat equation

In this project you will combine some of the theory and practical skills you learned in class to solve a real world problem. In addition to writing code that solves the program, you will use the following software practices:

- use source code control,
- give the program checkpoint/restart capability, and
- use basic profiling and visualization of the results.

The 1D Heat Equation (see section 4.3) is given by

$$\frac{\partial T(x,t)}{\partial t} = \alpha \frac{\partial^2 T(x,y)}{\partial x^2} + q(x,t)$$

where $t \geq 0$ and $x \in [0,1]$, subject to boundary conditions

$$T(x,0) = T_0(x), \quad \text{for } x \in [0,1],$$

$$T(0,t) = T_a(t),\ T(1,t) = T_b(t), \quad \text{for } t > 0.$$

You will solve this problem using the explicit and implicit Euler methods.

C.1.1 Software

Write your software using the PETSc library (see tutorial B.8. In particular, use the `MatMult` routine for matrix-vector multiplication and `KSPSolve` for linear system solution. Exception: code the Euler methods yourself.

Be sure to use a Makefile for building your project (tutorial B.5).

Add your source files, Makefile, and job scripts to an svn repository (tutorial B.6); do not add binaries or output files. Make sure the repository is accessible by the `istc00` account (which is in the same Unix group as your account) and that there is a README file with instructions on how to build and run your code.

Implement a checkpoint/restart facility by writing vector data, size of the time step, and other necessary items, to an `hdf5` file (tutorial B.7). Your program should be able to read this file and resume execution.

C.1.2 Tests

Do the following tests on a single core.

Method stability

Run your program with

$$\begin{cases} q = \sin \ell \pi x \\ T_0(x) = e^x \\ T_a(t) = T_b(t) = 0 \\ alpha = 1 \end{cases}$$

Take a space discretization $h = 10^{-2}$; try various time steps and show that the explicit method can diverge. What is the maximum time step for which it is stable?

For the implicit method, at first use a direct method to solve the nnsystem. This corresponds to PETSc options `KSPPREONLY` and `PCLU` (see section 5.5.10 and the PETSc tutorial, B.8).

Now use an iterative method (for instance `KSPCG` and `PCJACOBI`); is the method still stable? Explore using a low convergence tolerance and large time steps.

Since the forcing function q and the boundary conditions have no time dependence, the solution $u(\cdot, t)$ will converge to a *steady state* solution $u_\infty(x)$ as $t \to \infty$. What is the influence of the time step on the speed with which implicit method converges to this steady state?

Run these tests with various values for ℓ.

Timing

If you run your code with the commandline option `-log_summary`, you will get a table of timings of the various PETSc routines. Use that to do the following timing experiments.

- Construct your coefficient matrix as a dense matrix, rather than sparse. Report on the difference in total memory, and the runtime and flop rate of doing one time step. Do this for both the explicit and implicit method and explain the results.
- With a sparse coefficient matrix, report on the timing of a single time step. Discuss the respective flop counts and the resulting performance.

Restart

Implement a restart facility: every 10 iterations write out the values of the current iterate, together with values of Δx, Δt, and ℓ. Add a flag `-restart` to your program that causes it to read the restart file and resume execution, reading all parameters from the restart file.

Run your program for 25 iterations, and restart, causing it to run again from iteration 20. Check that the values in iterations $20 \ldots 25$ match.

C.1.3 Parallelism

Do the following tests, running your code with up to 16 cores.

- At first test the explicit method, which should be perfectly parallel. Report on actual speedup attained. Try larger and smaller problem sizes and report on the influence of the problem size.
- The above settings for the implicit method (`KSPPREONLY` and `PCLU`) lead to a runtime error. Instead, let the system be solved by an iterative method. Read the PETSc manual and web pages to find out some choices of iterative method and preconditioner and try them. Report on their efficacy.

C.1.4 Reporting

Write your report using LaTeX (tutorial B.2). Use both tables and graphs to report numerical results. Use `gnuplot` (tutorial B.9) or a related utility for graphs.

Appendix D

Codes

This section contains several simple codes that illustrate various issues relating to the performance of a single CPU. The explanations can be found in section 1.5.

D.1 Hardware event counting

The codes in this chapter make calls to a library named PAPI for 'Performance Application Programming Interface' [17, 5]. This is a portable set of calls to query the hardware counters that are built into most processors. Since these counters are part of the processor hardware, they can measure detailed events such as cache misses without this measurement process disturbing the phenomenon it is supposed to observe.

While using hardware counters is fairly straightforward, the question of whether what they are reporting is what you actually meant to measure is another matter altogether. For instance, the presence of hardware prefetch streams (section 1.2.5) implies that data can be loaded into cache without this load being triggered by a cache miss. Thus, the counters may report numbers that seem off, or even impossible, under a naive interpretation.

D.2 Cache size

This code demonstrates the fact that operations are more efficient if data is found in L1 cache, than in L2, L3, or main memory. To make sure we do not measure any unintended data movement, we perform one iteration to bring data in the cache before we start the timers.

```
/*
 * File:    size.c
 * Author:  Victor Eijkhout <eijkhout@tacc.utexas.edu>
 *
 * Usage: size
 */

#include "papi_test.h"
```

```
extern int TESTS_QUIET;         /* Declared in test_utils.c */

#define PCHECK(e) \
  if (e!=PAPI_OK) \
    {printf("Problem in papi call, line %d\n",__LINE__); return 1;}
#define NEVENTS 3
#define NRUNS 200
#define L1WORDS 8096
#define L2WORDS 100000

int main(int argc, char **argv)
{
  int events[NEVENTS] =
    {
      PAPI_TOT_CYC,/* total cycles */
      PAPI_L1_DCM, /* stalls on L1 cache miss */
      PAPI_L2_DCM, /* stalls on L2 cache miss */
    };
  long_long values[NEVENTS];
  PAPI_event_info_t info, info1;
  const PAPI_hw_info_t *hwinfo = NULL;
  int retval,event_code, m,n, i,j,size, arraysize;
  const PAPI_substrate_info_t *s = NULL;
  double *array;

   tests_quiet(argc, argv);     /* Set TESTS_QUIET variable */
   retval = PAPI_library_init(PAPI_VER_CURRENT);
   if (retval != PAPI_VER_CURRENT)
      test_fail(__FILE__, __LINE__, "PAPI_library_init", retval);
   {
     int i;
     for (i=0; i<NEVENTS; i++) {
       retval = PAPI_query_event(events[i]) ; PCHECK(retval);
     }
   }

   /* declare an array that is more than twice the L2 cache size */
   arraysize=2*L2WORDS;
   array = (double*) malloc(arraysize*sizeof(double));

   for (size=L1WORDS/4; size<arraysize; size+=L1WORDS/4) {
     printf("Run: data set size=%d\n",size);

     /* clear the cache by dragging the whole array through it */
     for (n=0; n<arraysize; n++) array[n] = 0.;
     /* now load the data in the highest cache level that fits */
```

```
    for (n=0; n<size; n++) array[n] = 0.;

    retval = PAPI_start_counters(events,NEVENTS); PCHECK(retval);
    /* run the experiment */
    for (i=0; i<NRUNS; i++) {
      for (j=0; j<size; j++) array[j] = 2.3*array[j]+1.2;
    }
    retval = PAPI_stop_counters(values,NEVENTS); PCHECK(retval);
    printf("size=%d\nTot cycles: %d\n",size,values[0]);
    printf("cycles per array loc: %9.5f\n",size,values[0]/(1.*NRUNS*size));
    printf("L1 misses:\t%d\nfraction of L1 lines missed:\t%9.5f\n",
   values[1],values[1]/(size/8.));
    printf("L2 misses:\t%d\nfraction of L2 lines missed:\t%9.5f\n",
   values[2],values[2]/(size/8.));
    printf("\n");
  }
  free(array);

  return 0;
}
```

D.3 Cachelines

This code illustrates the need for small strides in vector code. The main loop operates on a vector, progressing by a constant stride. As the stride increases, runtime will increase, since the number of cachelines transferred increases, and the bandwidth is the dominant cost of the computation.

There are some subtleties to this code: in order to prevent accidental reuse of data in cache, the computation is preceded by a loop that accesses at least twice as much data as will fit in cache. As a result, the array is guaranteed not to be in cache.

```
/*
 * File:    line.c
 * Author:  Victor Eijkhout <eijkhout@tacc.utexas.edu>
 *
 * Usage: line
 */

#include "papi_test.h"
extern int TESTS_QUIET;         /* Declared in test_utils.c */

#define PCHECK(e) \
  if (e!=PAPI_OK) \
    {printf("Problem in papi call, line %d\n",__LINE__); return 1;}
#define NEVENTS 4
```

```
#define MAXN 10000
#define L1WORDS 8096
#define MAXSTRIDE 16

int main(int argc, char **argv)
{
  int events[NEVENTS] =
    {PAPI_L1_DCM,/* stalls on L1 cache miss */
     PAPI_TOT_CYC,/* total cycles */
     PAPI_L1_DCA, /* cache accesses */
     1073872914 /* L1 refills */};
  long_long values[NEVENTS];
  PAPI_event_info_t info, info1;
  const PAPI_hw_info_t *hwinfo = NULL;
  int retval,event_code, m,n, i,j,stride, arraysize;
  const PAPI_substrate_info_t *s = NULL;
  double *array;

   tests_quiet(argc, argv);     /* Set TESTS_QUIET variable */
   retval = PAPI_library_init(PAPI_VER_CURRENT);
   if (retval != PAPI_VER_CURRENT)
      test_fail(__FILE__, __LINE__, "PAPI_library_init", retval);
   {
     int i;
     for (i=0; i<NEVENTS; i++) {
       retval = PAPI_query_event(events[i]) ; PCHECK(retval);
     }
   }

   /* declare an array that is more than twice the cache size */
   arraysize=2*L1WORDS*MAXSTRIDE;
   array = (double*) malloc(arraysize*sizeof(double));

   for (stride=1; stride<=MAXSTRIDE; stride++) {
     printf("Run: stride=%d\n",stride);
     /* clear the cache by dragging the whole array through it */
     for (n=0; n<arraysize; n++) array[n] = 0.;

     retval = PAPI_start_counters(events,NEVENTS); PCHECK(retval);
     /* run the experiment */
     for (i=0,n=0; i<L1WORDS; i++,n+=stride) array[n] = 2.3*array[n]+1.2;
     retval = PAPI_stop_counters(values,NEVENTS); PCHECK(retval);
     printf("stride=%d\nTot cycles: %d\n",stride,values[1]);
     printf("L1 misses:\t%d\naccesses per miss:\t%9.5f\n",
    values[0],(1.*L1WORDS)/values[0]);
     printf("L1 refills:\t%d\naccesses per refill:\t%9.5f\n",
```

```
     values[3],(1.*L1WORDS)/values[3]);
      printf("L1 accesses:\t%d\naccesses per operation:\t%9.5f\n",
     values[2],(1.*L1WORDS)/values[2]);
      printf("\n");
    }
    free(array);

    return 0;
}
```

Note that figure 1.6 in section 1.5.3 only plots up to stride 8, while the code computes to 16. In fact, at stride 12 the prefetch behaviour of the Opteron changes, leading to peculiarities in the timing, as shown in figure D.1.

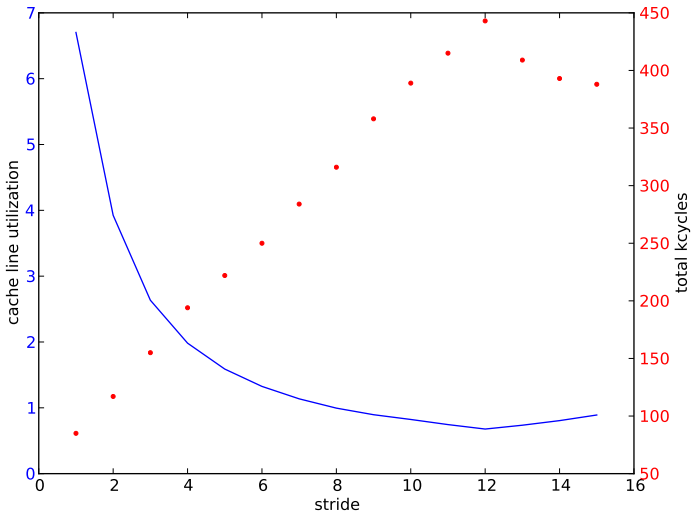


Figure D.1: Run time in kcycles and L1 reuse as a function of stride

D.4 Cache associativity

This code illustrates the effects of cache associativity;see sections 1.2.4.6 and 1.5.5 for a detailed explanation. A number of vectors (dependent on the inner loop variable `i`) is traversed simultaneously. Their lengths are chosen to induce cache conflicts. If the number of vectors is low enough, cache associativity will resolve these conflicts; for higher values of `m` the runtime will quickly increase. By allocating the vectors with a larger size, the cache conflicts go away.

```
/*
 * File:    assoc.c
```

```
 * Author:  Victor Eijkhout <eijkhout@tacc.utexas.edu>
 *
 * Usage: assoc m n
 */

#include "papi_test.h"
extern int TESTS_QUIET;         /* Declared in test_utils.c */

#define PCHECK(e) if (e!=PAPI_OK) {printf("Problem in papi call, line %d\n",__LINE__); ret
#define NEVENTS 2
#define MAXN 20000

/* we are assuming array storage in C row mode */
#if defined(SHIFT)
#define INDEX(i,j,m,n) (i)*(n+8)+(j)
#else
#define INDEX(i,j,m,n) (i)*(n)+(j)
#endif

int main(int argc, char **argv)
{
   int events[NEVENTS] = {PAPI_L1_DCM,PAPI_TOT_CYC}; long_long values[NEVENTS];
   PAPI_event_info_t info, info1;
   const PAPI_hw_info_t *hwinfo = NULL;
   int retval,event_code, m,n, i,j;
   const PAPI_substrate_info_t *s = NULL;
   double *array;

   tests_quiet(argc, argv);     /* Set TESTS_QUIET variable */
   retval = PAPI_library_init(PAPI_VER_CURRENT);
   if (retval != PAPI_VER_CURRENT)
      test_fail(__FILE__, __LINE__, "PAPI_library_init", retval);
   {
     int i;
     for (i=0; i<NEVENTS; i++) {
       retval = PAPI_query_event(events[i]); PCHECK(retval);
     }
   }
   /*
   if (argc<3) {
     printf("Usage: assoc m n\n"); return 1;
   } else {
     m = atoi(argv[1]); n = atoi(argv[2]);
   } printf("m,n = %d,%d\n",m,n);
   */
```

```
#if defined(SHIFT)
   array = (double*) malloc(13*(MAXN+8)*sizeof(double));
#else
   array = (double*) malloc(13*MAXN*sizeof(double));
#endif

   /* clear the array and bring in cache if possible */
   for (m=1; m<12; m++) {
     for (n=2048; n<MAXN; n=2*n) {
       printf("Run: %d,%d\n",m,n);
#if defined(SHIFT)
       printf("shifted\n");
#endif

   for (i=0; i<=m; i++)
     for (j=0; j<n; j++)
       array[INDEX(i,j,m+1,n)] = 0.;

   /* access the rows in a way to cause cache conflicts */
   retval = PAPI_start_counters(events,NEVENTS); PCHECK(retval);
   for (j=0; j<n; j++)
     for (i=1; i<=m; i++)
       array[INDEX(0,j,m+1,n)] += array[INDEX(i,j,m+1,n)];
   retval = PAPI_stop_counters(values,NEVENTS); PCHECK(retval);
   printf("m,n=%d,%d\n#elements:\t%d\nTot cycles: %d\nL1 misses:\t%d\nmisses per accumulation:\t
  m,n,m*n,values[1],values[0],values[0]/(1.*n));

     }
   }
   free(array);

   return 0;
}
```

D.5 TLB

This code illustrates the behaviour of a *TLB*; see sections 1.2.7 and 1.5.4 for a thorough explanation. A two-dimensional array is declared in column-major ordering (Fortran style). This means that striding through the data by varying the i coordinate will have a high likelihood of TLB hits, since all elements on a page are accessed consecutively. The number of TLB entries accessed equals the number of elements divided by the page size. Striding through the array by the j coordinate will have each next element hitting a new page, so TLB misses will ensue when the number of columns is larger than the number of TLB entries.

```
/*
 * File:    tlb.c
```

```
 * Author:  Victor Eijkhout <eijkhout@tacc.utexas.edu>
 */

#include "papi_test.h"
extern int TESTS_QUIET;         /* Declared in test_utils.c */

#define PCHECK(e) if (e!=PAPI_OK) {printf("Problem in papi call, line %d\n",__LINE__); ret
#define NEVENTS 2
/* we are assuming array storage in Fortran column mode */
#define INDEX(i,j,m,n) i+j*m

double *array;

void clear_right(int m,int n) {
  int i,j;
  for (j=0; j<n; j++)
    for (i=0; i<m; i++)
      array[INDEX(i,j,m,n)] = 0;
  return;
}

void clear_wrong(int m,int n) {
  int i,j;
  for (i=0; i<m; i++)
    for (j=0; j<n; j++)
      array[INDEX(i,j,m,n)] = 0;
  return;
}

void do_operation_right(int m,int n) {
  int i,j;
  for (j=0; j<n; j++)
    for (i=0; i<m; i++)
      array[INDEX(i,j,m,n)] = array[INDEX(i,j,m,n)]+1;
  return;
}

void do_operation_wrong(int m,int n) {
  int i,j;
  for (i=0; i<m; i++)
    for (j=0; j<n; j++)
      array[INDEX(i,j,m,n)] = array[INDEX(i,j,m,n)]+1;
  return;
}

#define COL 1
```

```
#define ROW 2
int main(int argc, char **argv)
{
   int events[NEVENTS] = {PAPI_TLB_DM,PAPI_TOT_CYC}; long_long values[NEVENTS];
   int retval,order=COL;
   PAPI_event_info_t info, info1;
   const PAPI_hw_info_t *hwinfo = NULL;
   int event_code;
   const PAPI_substrate_info_t *s = NULL;

   tests_quiet(argc, argv);     /* Set TESTS_QUIET variable */
   if (argc==2 && !strcmp(argv[1],"row")) {
     printf("wrong way\n"); order=ROW;
   } else printf("right way\n");

   retval = PAPI_library_init(PAPI_VER_CURRENT);
   if (retval != PAPI_VER_CURRENT)
      test_fail(__FILE__, __LINE__, "PAPI_library_init", retval);

   {
     int i;
     for (i=0; i<NEVENTS; i++) {
       retval = PAPI_query_event(events[i]); PCHECK(retval);
     }
   }

#define M 1000
#define N 2000
   {
     int m,n;
     m = M;
     array = (double*) malloc(M*N*sizeof(double));
     for (n=10; n<N; n+=10) {
       if (order==COL)
 clear_right(m,n);
       else
 clear_wrong(m,n);
       retval = PAPI_start_counters(events,NEVENTS); PCHECK(retval);
       if (order==COL)
 do_operation_right(m,n);
       else
 do_operation_wrong(m,n);
       retval = PAPI_stop_counters(values,NEVENTS); PCHECK(retval);
       printf("m,n=%d,%d\n#elements:\t%d\nTot cycles: %d\nTLB misses:\t%d\nmisses per column:\t%
      m,n,m*n,values[1],values[0],values[0]/(1.*n));
     }
```

```
      free(array);
   }

   return 0;
}
```

Bibliography

[1] Chapel programming language homepage. http://chapel.cray.com/.

[2] Development of ternary computers at Moscow State University, note = http://www.computer-museum.ru/english/setun.htm.

[3] Ieee 754: Standard for binary floating-point arithmetic. http://grouper.ieee.org/groups/754.

[4] Kendall square research. http://en.wikipedia.org/wiki/Kendall_Square_Research.

[5] Performance application programming interface. http://icl.cs.utk.edu/papi/.

[6] Project fortress homepage. http://projectfortress.sun.com/Projects/Community.

[7] Setun. http://en.wikipedia.org/wiki/Setun.

[8] Ternary computer. http://en.wikipedia.org/wiki/Ternary_computer.

[9] Texas advanced computing center: Sun constellation cluster: Ranger. http://www.tacc.utexas.edu/resources/hpc/constellation.

[10] Universal parallel c at george washingon university. http://upc.gwu.edu/.

[11] G. Amdahl. The validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the AFIPS Computing Conference*, volume 30, pages 483–485, 1967.

[12] O. Axelsson and A.V. Barker. *Finite element solution of boundary value problems. Theory and computation.* Academic Press, Orlando, Fl., 1984.

[13] K.E. Batcher. MPP: A high speed image processor. In *Algorithmically Specialized Parallel Computers.* Academic Press, New York, 1985.

[14] Abraham Berman and Robert J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* SIAM, 1994. originally published by Academic Press, 1979, New York.

[15] Netlib.org BLAS reference implementation. http://www.netlib.org/blas.

[16] BOOST interval arithmetic library. http://www.boost.org/libs/numeric/interval/doc/interval.htm.

[17] S. Browne, J Dongarra, N. Garner, G. Ho, and P. Mucci. A portable programming interface for performance evaluation on modern processors. *International Journal of High Performance Computing Applications*, 14:189–204, Fall 2000.

[18] Ernie Chan, Marcel Heimlich, Avi Purkayastha, and Robert van de Geijn. Collective communication: theory, practice, and experience. *Concurrency and Computation: Practice and Experience*, 19:1749–1783, 2007.

[19] Barbara Chapman, Gabriele Jost, and Ruud van der Pas. *Using OpenMP: Portable Shared Memory Parallel Programming*, volume 10 of *Scientific Computation Series.* MIT Press, ISBN 0262533022, 9780262533027, 2008.

[20] Yaeyoung Choi, Jack J. Dongarra, Roldan Pozo, and David W. Walker. Scalapack: a scalable linear algebra library for distributed memory concurrent computers. In *Proceedings of the fourth symposium on the frontiers of massively parallel computation (Frontiers '92), McLean, Virginia, Oct 19–21, 1992*, pages 120–127, 1992.

[21] Charles Clos. A study of non-blocking switching networks. *Bell System Technical Journal*, 32:406–242, 1953.

[22] Ashish Deshpande and Martin Schultz. Efficient parallel programming with linda. In *In Supercomputing '92 Proceedings*, pages 238–244, 1992.

[23] E. W. Dijkstra. Cooperating sequential processes. `http://www.cs.utexas.edu/users/EWD/transcriptions/EWD01xx/EWD123.html`. Technological University, Eindhoven, The Netherlands, September 1965.

[24] Edsger W. Dijkstra. Programming as a discipline of mathematical nature. *Am. Math. Monthly*, 81:608–612, 1974.

[25] J. Dongarra, A. Geist, R. Manchek, and V. Sunderam. Integrated PVM Framework Supports Heterogeneous Network Computing. *Computers in Physics*, 7(2):166–75, April 1993.

[26] J. J. Dongarra. *The LINPACK benchmark: An explanation*, volume 297, chapter Supercomputing 1987, pages 456–474. Springer-Verlag, Berlin, 1988.

[27] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Iain Duff. A set of level 3 basic linear algebra subprograms. *ACM Transactions on Mathematical Software*, 16(1):1–17, March 1990.

[28] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Richard J. Hanson. An extended set of FORTRAN basic linear algebra subprograms. *ACM Transactions on Mathematical Software*, 14(1):1–17, March 1988.

[29] Dr. Dobbs. Complex arithmetic: in the intersection of C and C++. `http://www.ddj.com/cpp/184401628`.

[30] V. Faber and T. A. Manteuffel. Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. Numer. Anal.*, 21(2):352–362, 1984.

[31] R.D. Falgout, J.E. Jones, and U.M. Yang. Pursuing scalability for hypre's conceptual interfaces. Technical Report UCRL-JRNL-205407, Lawrence Livermore National Lab, 2004. submitted to ACM Transactions on Mathematical Software.

[32] D. C. Fisher. Your favorite parallel algorithms might not be as fast as you think. *IEEE Trans. Computers*, 37:211–213, 1988.

[33] M. Flynn. Some computer organizations and their effectiveness. *IEEE Trans. Comput.*, C-21:948, 1972.

[34] D. Frenkel and B. Smit. Understanding molecular simulations: From algorithms to applications, 2nd edition. 2002.

[35] Roland W. Freund and Noël M. Nachtigal. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numer. Math.*, 60:315–339, 1991.

[36] A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, and V. Sunderam. *PVM: A Users' Guide and Tutorial for Networked Parallel Computing*. MIT Press, 1994. The book is available electronically, the url is `ftp://www.netlib.org/pvm3/book/pvm-book.ps`.

[37] David Gelernter and Nicholas Carriero. Coordination languages and their significance. *Commun. ACM*, 35(2):97–107, 1992.

[38] GNU multiple precision library. `http://gmplib.org/`.

[39] Dominik Göddeke, Robert Strzodka, Jamaludin Mohd-yusof, Patrick Mccormick, Hilmar Wobker, Christian Becker, and Stefan Turek. Using gpus to improve multigrid solver performance on a cluster, accepted for publication. *International Journal of Computational Science and Engineering*, 4:36–55, 2008.

[40] Stefan Goedecker and Adolfy Hoisie. *Performance Optimization of Numerically Intensive Codes*. SIAM, 2001.

[41] David Goldberg. What every computer scientist should know about floating-point arithmetic. *Computing Surveys*, March 1991.

[42] G. H. Golub and D. P. O'Leary. Some history of the conjugate gradient and Lanczos algorithms: 1948-1976. 31:50–102, 1989.

[43] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, second edition edition, 1989.

[44] F. Gray. Pulse code communication. U.S. Patent 2,632,058, March 17, 1953 (filed Nov. 1947).

[45] Ronald I. Greenberg and Charles E. Leiserson. Randomized routing on fat-trees. In *Advances in Computing Research*, pages 345–374. JAI Press, 1989.

[46] William Gropp, Steven Huss-Lederman, Andrew Lumsdaine, Ewing Lusk, Bill Nitzberg, William Saphir, and Marc Snir. *MPI: The Complete Reference, Volume 2 - The MPI-2 Extensions*. MIT Press, 1998.

[47] William Gropp, Thomas Sterling, and Ewing Lusk. *Beowulf Cluster Computing with Linux, 2nd Edition*. MIT Press, 2003.

[48] Don Heller. A survey of parallel algorithms in numerical linear algebra. *SIAM Review*, 20:740–777, 1978.

[49] John L. Hennessy and David A. Patterson. *Computer Architecture, A Quantitative Approach*. Morgan Kaufman Publishers, 3rd edition edition, 1990, 3rd edition 2003.

[50] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Nat. Bur. Stand. J. Res.*, 49:409–436, 1952.

[51] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.

[52] C. A. R. Hoare. An axiomatic basis for computer programming. *Communications of the ACM*, 12:576–580, 1969.

[53] Interval arithmetic. `http://en.wikipedia.org/wiki/Interval_(mathematics)`.

[54] C.R. Jesshope and R.W. Hockney editors. The DAP approach, volume 2. pages 311–329. Infotech Intl. Ltd., Maidenhead, 1979.

[55] Michael Karbo. PC architecture. `http://www.karbosguide.com/books/pcarchitecture/chapter00.htm`.

[56] Helmut Kopka and Patrick W. Daly. *A Guide to LaTeX*. Addison-Wesley, first published 1992.

[57] Milind Kulkarni, Martin Burtscher, Rajasekhar Inkulu, Keshav Pingali, and Calin Cascaval. How much parallelism is there in irregular applications? In *Principles and Practices of Parallel Programming (PPoPP)*, 2009.

[58] L. Lamport. *LaTeX, a Document Preparation System*. Addison-Wesley, 1986.

[59] C. Lanczos. Solution of systems of linear equations by minimized iterations. *Journal of Research, Nat. Bu. Stand.*, 49:33–53, 1952.

[60] Rubin H Landau, Manual José Páez, and Cristian C. Bordeianu. *A Survey of Computational Physics*. Princeton University Press, 2008.

[61] C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh. Basic linear algebra subprograms for Fortran usage. *ACM Transactions on Mathematical Software*, 5(3):308–323, September 1979.

[62] Charles E. Leiserson. Fat-Trees: Universal networks for hardware-efficient supercomputing. *IEEE Trans. Comput*, C-34:892–901, 1985.

[63] Robert Mecklenburg. *Managing Projects with GNU Make*. O'Reilly Media, 3rd edition edition, 2004. Print ISBN:978-0-596-00610-5 ISBN 10:0-596-00610-1 Ebook ISBN:978-0-596-10445-0 ISBN 10:0-596-10445-6.

[64] Frank Mittelbach, Michel Goossens, Johannes Braams, David Carlisle, and Chris Rowley. *The LaTeX Companion, 2nd edition*. Addison-Wesley, 2004.

[65] Tobi Oetiker. The not so short introductino to LaTeX. `http://tobi.oetiker.ch/lshort/`.

[66] National Institute of Standards and Technology. Matrix market. `http://math.nist.gov/MatrixMarket`.

[67] S. Otto, J. Dongarra, S. Hess-Lederman, M. Snir, and D. Walker. *Message Passing Interface: The Complete Reference*. The MIT Press, 1995.

[68] Michael L. Overton. *Numerical Computing with IEEE Floating Point Arithmetic*. SIAM, Philadelphia PA, 2001.

[69] V. Ya. Pan. New combination sof methods for the acceleration of matrix multiplication. *Comp. & Maths. with Appls.*, 7:73–125, 1981.

[70] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.*, 117:1–19, 1995.

[71] M. Püschel, B. Singer, J. Xiong, J. M. F. Moura, J. Johnson, D. Padua, M. Veloso, and R. W. Johnson. SPIRAL: A generator for platform-adapted libraries of signal processing algorithms. *Int'l Journal of High Performance Computing Applications*, 18(1):21–45, 2004.

[72] J.K. Reid. On the method of conjugate gradients for the solution of large sparse systems of linear equations. In J.K. Reid, editor, *Large sparse sets of linear equations*, pages 231–254. Academic Press, London, 1971.

[73] Tetsuya Sato. The earth simulator: Roles and impacts. *Nuclear Physics B - Proceedings Supplements*, 129-130:102 – 108, 2004. Lattice 2003.

[74] D. E. Shaw. A fast, scalable method for the parallel evaluation of distance-limited pairwise particle interactions. *J. Comput. Chem.*, 26:1318–1328, 2005.

[75] Marc Snir, Steve Otto, Steven Huss-Lederman, David Walker, and Jack Dongarra. *MPI: The Complete Reference, Volume 1, The MPI-1 Core*. MIT Press, second edition edition, 1998.

[76] V. Strassen. Gaussian elimination is not optimal. *Numer. Math.*, 13:354–356, 1969.

[77] TeX frequently asked questions.

[78] R. van de Geijn, Philip Alpatov, Greg Baker, Almadena Chtchelkanova, Joe Eaton, Carter Edwards, Murthy Guddati, John Gunnels, Sam Guyer, Ken Klimkowski, Calvin Lin, Greg Morrow, Peter Nagel, James Overfelt, and Michelle Pal. Parallel linear algebra package (PLAPACK): Release r0.1 (beta) users' guide. 1996.

[79] Robert A. van de Geijn. *Using PLAPACK: Parallel Linear Algebra Package*. The MIT Press, 1997.

[80] Robert A. van de Geijn and Enrique S. Quintana-Ortí. *The Science of Programming Matrix Computations*. `www.lulu.com`, 2008.

[81] Henk van der Vorst. Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 13:631–644, 1992.

[82] Richard S. Varga. *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1962.

[83] R. Clint Whaley, Antoine Petitet, and Jack J. Dongarra. Automated empirical optimization of software and the ATLAS project. *Parallel Computing*, 27(1–2):3–35, 2001. Also available as University of Tennessee LAPACK Working Note #147, UT-CS-00-448, 2000 (`www.netlib.org/lapack/lawns/lawn147.ps`).

[84] O. Widlund. On the use of fast methods for separable finite difference equations for the solution of general elliptic problems. In D.J. Rose and R.A. Willoughby, editors, *Sparse matrices and their applications*, pages 121–134. Plenum Press, New York, 1972.

[85] Barry Wilkinson and Michael Allen. *Parallel Programming*. Prentice Hall, New Jersey, 1999.

[86] J.H. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice-Hall, Englewood Cliffs, N.J., 1963.

[87] Randy Yates. Fixed point: an introduction. `http://www.digitalsignallabs.com/fp.pdf`, 2007.

[88] David M. Young. *Iterative method for solving partial differential equations of elliptic type*. PhD thesis, Harvard Univ., Cambridge, MA, 1950.

Appendix E

Index and list of acronyms

BEM Boundary Element Method
BLAS Basic Linear Algebra Subprograms
BVP Boundary Value Problem
CAF Co-array Fortran
CCS Compressed Column Storage
CG Conjugate Gradients
CGS Classical Gram-Schmidt
COO Coordinate Storage
CRS Compressed Row Storage
CSV comma-separated values
DAG Directed Acyclic Graph
FD Finite Difference
FEM Finite Element Method
FFT Fast Fourier Transform
FOM Full Orthogonalization Method
FSA Finite State Automaton
FSB Front-Side Bus
GMRES Generalized Minimum Residual
GPGPU General Purpose Graphics Processing Unit
GPU Graphics Processing Unit
GS Gram-Schmidt
HPC High-Performance Computing
HPFHigh Performance Fortran
IBVP Initial Boundary Value Problem
ILP Instruction Level Parallelism
ILU Incomplete LU
IVP Initial Value Problem
LAN Local Area Network
LRU Least Recently Used
MGS Modiefed Gram-Schmidt
MIMD Multiple Instruction Multiple Data
MPI Message Passing Interface
NUMA Non-Uniform Memory Access
ODE Ordinary Diffential Equation
OS Operating System
PGAS Partitioned Global Address Space
PDE Partial Diffential Equation
PRAM Programmable Random Access Machine
SAN Storage Area Network
SIMD Single Instruction Multiple Data
SM Streaming Multiprocessor
SMP Symmetric Multi Processor
SOR Successive Over-Relaxation
SP Streaming Processor
SPMD Single Program Multiple Data
SPD symmetric positive definite
TLB Translation Look-aside Buffer
UMA Uniform Memory Access
UPC Unified Parallel C
WAN Wide Area Network

Index